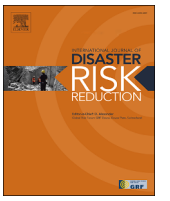
International Journal of Disaster Risk Reduction 86 (2023) 103502

Contents lists available at ScienceDirect

International Journal of Disaster Risk Reduction

journal homepage: www.elsevier.com/locate/ijdrr



A quantification of the reliability of self-reports following a simulated stressful event

Alastair Shipman [a, *], Guillaume Dezecache [b], Arnab Majumdar [a]

[a] Civil and Environmental Engineering, Imperial College London, UK
[b] Universite Clermont Auvergne, CNRS, LAPSCO, Clermont-Ferrand, France

ABSTRACT

Interviews and surveys are the most commonly used data-gathering and data-generating techniques when investigating human behaviour in emergencies. However, these approaches suffer from several limitations, including potential errors in memory accuracy, a lack of quantitative reliability. This study focuses on a survey performed on participants who had taken part in a stressful experiment. The survey was carried out three months afterwards, asking them to recall their experience. Analysis of this data quantitatively assesses their recall, across multiple different domains. This study observed several differences between experimental and control group participants, as well as differences between participants in VR and Physical experimental groups. However, it observes no increase in confabulation as a result of increased stress. The outcome of this study is to provide insight into the quantitative reliability of interviews and surveys of people involved in emergencies.

1. Introduction

According the Global Terrorism Database [1] the past decade has seen a rise in the number of terrorist attacks throughout the world. The methods employed by terrorists have changed as well, with the rise of attacks that require less preparation, are much harder to detect and prevent, and involve the use of diverse materials such as cars and knives [2]. With the rise of this form of terrorism, urgent questions have emerged in the field of psychology and the social sciences, including how to best mitigate and Control the outcome of these situations. Control measures such as communication with the public, understanding their likely responses and promoting swift evacuation during a terrorist attack all have the potential to save lives. Yet, this question requires resolving another interrogation: that of understanding the behaviour of individuals and groups when collectively confronted by a terrorist scenario. This problem is critical to inform the delivery of appropriate emergency guidelines [3], or whenever possible, to best guide architectural choices in designing buildings [4]. While as of yet the majority of research in these areas has been devoted to people's reactions during natural disasters [5], and during fire evacuations [6], recent years have seen an increase of interest for works covering terrorist attacks [7].

Regarding social behaviour, it is often assumed that people invest in antisocial strategies when confronted to deadly danger [3,8–10]. Self-protective strategies maximize one own's immediate and short-term welfare, possibly at the expense of others [8]. They include non-social (fleeing) as well as social (pushing, trampling) behaviours. At odds with this 'competitive' view of human behaviour in emergency contexts, allo-protective behaviour is commonly reported in situations as diverse as boat sinking, stadium stampedes, bus bombing or mass shootings [5,11–13]. This suggests that cooperative tendencies are promoted, or at least maintained, in the face of deadly danger.

A primary issue with these conclusions is that, while there are examples of objective data sources such as CCTV [14,14–16], these empirical findings that are almost uniformly obtained through interview reports of self- and other-behaviours. It remains the case that there are very few sources of objective data such as video footage or experimental approaches [10].

* Corresponding author.
E-mail address: as15516@ic.ac.uk (A. Shipman).

https://doi.org/10.1016/j.ijdrr.2022.103502
Received 7 February 2022; Received in revised form 1 December 2022; Accepted 15 December 2022
Available online 23 December 2022

Such measures, while they may be the most widespread way of getting access to insights into life-threatening situations are prone to a number of biases such as leading questions [17], social desirability, and maintenance of a positive image of oneself [18] and one's own group [19].

Another important, but poorly understood bias, revolves around memory- and recall-related issues. These biases, which are very often explicitly acknowledged, are such that they may underestimate or exaggerate the occurrence of specific behaviours (e.g. cooperation) during exposure to stressful stimuli, misrepresenting their true frequency. Acknowledging the existence of such biases does not mean we cannot trust the models formed by information gathered in testimonies of such scenarios, but rather that we should be able to, eventually, possess a theory of how, and to what extent, individual testimonies diverge from what actually occurred (the 'ground truth'). This understanding is required to best evaluate reports when building theories of how people behave in emergency situations, with respect to their social but also non-social behaviour, including the way they feel at the very moment of the attacks. Without this understanding, we cannot ascertain the disparity between an interviewee's transcript and the ground truth. Methodological principles have been developed to avoid some of the biases associated with self-reports (e.g., triangulation to ensure independent sources converge on reported facts), biases stemming from the reliability of testimonies (i.e., whether there is reasonable correspondence between what happened and what is reported) are more difficult to accommodate.

While there is extensive previous work in the area of self-reports within emergencies, including recalled experiences of floods [20], droughts [21], and other natural disasters [22] the literature on the quality of an individual's recollection of the details of threatening events (including one's own emotional and perceptual states) is vast and complex, and inconclusive at this stage. The current view is that memory is a constructive process, which can be particularly susceptible to the delivery of new information [23]. During exposure to a stressor, memory for specific and contextual details [24–26] can be impaired, making eyewitness reports probably unreliable, and at least highly questionable [27,28]. Further details can be found in Refs. [29,30] for a more comprehensive picture. In contrast to this, previous research has also shown that exposure to a stressor can highlight and improve the recall of elements of an individual's memory [31]. This dichotomy of results is unexplained.

Memories can suffer from further distortion when we consider the time between the stressor and the reports [32], as memories undergo natural decay [33] and time generally makes people more likely to encounter external information which can degrade their recall quality [32,34]. If the stressor is sufficiently intense, major psychiatric symptoms can appear, that may also affect recall of the event [35,36], particularly with respect to its temporality and one's own role in it [37]. Of particular interest here is the recall of the individual's and others' social behaviours, as well as the individual's emotional states and thoughts at the time of the event, and their relative accuracy when compared to the recall of other items. Although much controversy exists over the reliability of testimonies for a diversity of events, focus on 'attack-like' scenarios (notably mass shooting and stabbing attacks) and testimonies related to behaviours is much needed to progress on the understanding of the dynamics of social behaviour under stressful circumstances.

This study sought to evaluate the reliability of reports about a stressful event, including a comparative quantification of recall across multiple different domains. As part of this, participants took part in a study otherwise meant to investigate movements and trajectories when exposed to a simulated stressful event: the intrusion of an aggressive character, who is then chasing each of the participants to catch them and cause them to lose their participation payment. The authors collected a number of self-report and physiological measures prior and immediately after participation. The participants were then surveyed approximately three months afterwards, asking for their recall of what happened, their self-reported emotional states, their behaviour, the behaviour of others, as well as their recollection of the environment. This study then investigated the reliability of their reports, by quantifying the differences between the participants' recollection and the ground truth. To achieve this, the authors split the sets of questions into several different domains, and obtained a score for each participant within these domains. The authors quantified recall quality by assessing deviation from ground truth, with a score of 0 indicating perfect recall.

This paper deals with several groups of participants. The first group is the set of participants who took part in the physical version of the experiment (Physical); that is, in a real environment. The second group is the set of participants who took part in the virtual reality version of the experiment (VR), in a setting mimicking the real environment. These two groups were combined into a third group of participants who took part in a stressful experiment (Stress). This group contrasts with the final group, a subset of the group of participants from the Physical experiment, who took part in a set of Control experiments, without any stressing agent (Control).

The following section covers the methods used as part of this survey, including a description of the participants, the questionnaire procedure and the data coding. The results section details the results of the survey, comparing VR and Physical participants, as well as comparing Stress and Control participants. This section then continues to investigate the change in self-reported emotional state before then detailing regression models to understand the important predictor factors in recall quality, and comparing the level of confabulation between groups. Finally this paper then discusses the results, highlighting the most interesting results.

2. Methods

2.1. Ethics

This study was approved by the Ethics committee of the Imperial College London (SETREC reference numbers: 18IC4637, 19IC5216).

2.2. Participants

In total, 80 staff and student participants were recruited from Imperial College London to take part in the Physical experiment participants (26 females; 25.0 years old ± 3.88 SD. Of these, 19 were randomly allocated to the Control Branch, and the remaining 61 were allocated to the Experimental Branch. Additionally, 55 VR experiment participants (18 females; 21.3 years old ± 2.89 SD) were

recruited to take part in experiments that investigated behavioural responses to hostile aggressors (real confederate and actor in the Physical version; virtual actor in the VR version).

2.3. Summary of the pre-survey procedure

This section briefly describes the experimental procedures that generated the stressful stimuli. For further information on this, please see Ref. [38]).

2.3.1. Physical experiment

The physical version of the experiment took place with ten separate groups of participants (mean size = 8 ± 1.79 SD). Each experiment took place in a large room, and participants were each equipped with a helmet containing a position-tracker. The participants were provided with limited information only (see below), with the specific details of this provision being determined by a pilot study. For more information on this approach please see Ref. [35].

The participants were invited to take part in an experiment investigating movement behaviour, and told that they had the ability to earn up to £40 in Amazon vouchers upon successful completion of the experiment. They were also told that if they failed the experiment, they would only earn £5. Finally, they were told that they would be provided with tasks of varying levels of stress. Initially the participants were asked to work on distractor problems that were presented on sheets attached to the walls, but were informed that this task had no bearing on their potential earnings. After 5 min, and without the participants knowing this would happen, an intruder (a professional actor) entered the room, shouting and aggressively harassing one of the participants (a confederate and further professional actor). The intruder informed all of the participants that if he managed to touch them, they would fail the experiment and lose their potential earnings. At this point he started to chase all of the remaining participants, and, upon catching an individual, the aggressor told them to sit down. During this period the participants were tracked, providing information on their movements in response to aggressive hostile actors. The experiment stopped when the last participant had been caught.

In addition to the position measurements, this experiment gathered several different types of measurement. These were questionnaires (short form STAI-T, short form STAI-S, short form PANAS) and saliva samples which were used for corticosteroid quantification (specifically, cortisol). The STAI-T was performed prior to the experiment, while the remainder of the surveys were performed immediately afterwards, providing a quantification of the subjective emotional state that resulted from the experimental procedure. Four separate saliva samples were obtained, including a pre-experiment baseline, and samples at 20, 40 and 60 min after the aggressor entrance. These saliva measures were then 1D-spline interpolated, and the Min-Max difference was quantified (see Supplementary Information).

2.3.2. VR experiment

The procedure was similar as in the physical experiment except there was no problem-solving task and that the nine other avatars were not controlled by other actual participants but instead by the computational game engine. To ensure the participants retained a sense of social interaction, they were told at the time that the other participants were in another room and that they would meet each other after the experiment. At the end of the experimental procedures, the experiments were informed that the avatars were computer controlled, and then asked to rank how much they believed that they were humans. The results are shown below in Supplementary Information and indicate that participants reasonably thought they were taking part in the experiment with others.

2.3.3. Post-experimental survey procedure

Participants were informed at the experiment that a follow-up survey would be sent to them but were not told of its nature. The surveys were sent to all participants (for both the Physical and VR experiments) by email, asking them to respond to the survey via a link to a Qualtrics survey. These emails were sent 12 weeks after participation, and reminders were sent at to participants who had not completed the survey up to three times, at weekly intervals. In total 103 individuals, comprising 57 (71%) of the Physical experiment sample, and 46 (84%) of the VR experiment sample, completed the survey. These participants then filled in the surveys after an average of 93 days (mean = 93.07 days; SD = 7.82; range = 81–120).

Participants were asked to respond by the best of their memory, and not use the information contained in email correspondence. Note that there were slight differences between the content of the survey for the Physical and VR participants. Both surveys can be accessed here (Physical survey: [39], VR survey [40]) The surveys consisted of five parts. In the first part, participants were asked to provide general information about their participation, specifically: what happened (measured via free recall), which instructions were given by the instructor (free recall), the date and time of the experiment, how many other participants were involved, and how many of the participants were female. In the second part, the participants were asked to describe their situation within the experimental area throughout the duration of the experiment: they first indicated, using a map of the room, where they think they were during the non-stressful phase of the experiment, and again when they initially noticed the aggressor. They were also asked a question about the characteristics of the experimental area, including asking which items were used to mark boundaries between several parts of the room. Finally, they were asked how they behaved when noticing the aggressor (free recall). In the third part, participants were asked questions about their recall of the intruder: including the gender, hair colour, the trajectory taken, and a summary of any instructions given. The fourth part of the survey concerned the participant's recollection of the self- and other reactions to the aggressor. As part of this, participants were first asked about the nature of participant's own reaction within 5s after discovering the intruder, with a mutually exclusive choice between the following options: ‘starting going towards the exit’, ‘started moving towards another participant’, ‘started confronting the aggressor’, ‘stood still’ or ‘none of the above’. Then, participants were asked to decide whether a series of statements were true or false (i.e., whether they had occurred or not in their own experimental session). Statements were about self-,

other and collective behaviours towards the aggressor (e.g., 'I *did not* verbally confront the aggressor', '*No one* verbally confronted the aggressor') or towards others (e.g., '*No one* was holding hands'). To avoid any potential bias, participants were presented with a 50-50 selection of negative questions (e.g.: 'I *did not* verbally confront the aggressor') and positive questions (e.g.: 'I verbally confronted the aggressor'). References to physical interactions were included owing to previous work that has observed physical contact as a prevalent group behaviour in response to a perceived threat [41] The fifth and last part of the survey asked participants to recall their responses to the questionnaires immediately after the experimental procedure. These were repeats of the short form STAI-S and short form PANAS questionnaires. The positive emotional metrics (PANAS-P) were included in the analysis for the following reason: it was observed in the initial experiments that the level of positive emotions were independent of the level of negative emotions [38]. This contributes to anecdotal evidence that collective threats can lead to positive emotions (especially in hindsight) of 'being together'.

2.4. Data analysis

The coding of survey elements was performed to assess the distance between participants' actual behaviour and their response to corresponding survey items. For qualitative questions this involved coding by two of the authors (Cohen's Kappa = 0.834). The authors focussed on a list of items for which correspondence to actual behaviour could clearly be established. The selected items were then allocated to a maximum of four separate recall domains, including an "Overall" domain which was used as a summary of participant recall quality. The number of questions in each domain is detailed below in Table 1.

2.5. Domain scores

To create the domain scores, the raw responses to all questions were recorded, and then normalised, leading to a score between 0 and 1 for each question. The scores were calculated using the following normalisation procedure:

Raw Score = *Deviation from truth* (*quantitative, or coded qualitative*)

$$Normalized\ score, Q_i = Abs\left(\frac{Raw\ score}{Max(Range\ \ (Raw\ scores), Max\,(Raw\ scores)}\right) \tag{1}$$

The scores were normalised to ensure that the responses to different questions could be combined into domain scores as well as maintaining the quantification of recall accuracy (i.e., a score of zero indicates perfect recall). The normalisation procedure used in this study was primarily aimed at only considering positive values, allowing insight into the strength of the quality of recall over several domains.

Each individual's domain scores were calculated as the arithmetic mean of all scores within that specific domain. The domain scores were initially calculated as the average absolute normalised score for participants in each domain, using the normalisation method shown in Equation (1).

However, there was the potential for participants to either underestimate (a negative score) or overestimate (a positive score) their subjective responses within the PANAS and STAI surveys. This was only considered within Domain 3: Self-subjective (see Table 1), where directional deviations in recall strength have important conceptual differences.

In order to investigate any such directional differences, each applicable question was then considered prior to taking an absolute value. This led to the normalised score being to between −1 and 1.

2.6. Source of recall alteration

A multiple regression was performed to best understand the factors driving the recall of information with memory recall as dependent variable. The independent variables were the different measures of emotional state, including the psychological measures: PANAS and STAI-S/T, and the maximum salivary cortisol difference measure. Further independent variables included demographic (age, exercise level, and gender), as well as a binary measure differentiating the VR and Physical experiments.

2.7. Confabulation

Finally, the level of confabulation (the creation of false memories) was quantified, by counting the number of times where a participant has stated that a false event had occurred in response to a binary 'yes or no' question.

Table 1
Number of items per domain in surveys for both the Physical and VR experiments.

Domain	Number of items (Physical)	Number of items (VR)
Overall	22	30
Descriptive	6	7
Self-Subjective	3	4
Self-Behaviour	6	9
Aggressor-related	13	15
Neighbour-related	9	10
Spatial	2	9

3. Results

In this paper the Mann Whitney *U* Test is used to investigate differences between group values. This test is used as many of the group subsets did not satisfy the requirements for a *t*-test, such as normality. The following significance codes were used: *p < 0.05, **p < 0.01, ***p < 0.001. Any result that was significant at 10% or less was referred to as 'marginally' significant, †p < 0.1.

Fig. 4 shows a graphical depiction of results of the surveys, across all paradigms. A full table of the results is provided in the supplementary material.

3.1. Differences between paradigms

Within the Overall domain with respect to raw scores, there was no statistical difference for recall between the Control and the Physical paradigm (Mann Whitney U-Tests: U = 274, p = 0.11). It was found that recall was better regarding neighbour-related information for stress against Control conditions (Mean domain score stress = 0.205; Control = 0.236, U = 226.5, p = 0.02). The other comparisons were not statistically significant (see Supplementary Information).

Comparisons of raw scores between the Physical and VR groups revealed a marginally significant difference Overall (U = 665, p = 0.06), with higher recall for VR. Information related to Descriptive (Mean VR = 0.189; Mean Physical = 0.269; U = 370, p < 0.01) and neighbours (Mean VR = 0.140; Mean Physical = 0.207; U = 587, p = 0.01) were also better recalled for VR versus Physical. The other comparisons were not statistically significant (see Supplementary Information and Fig. 1).

3.2. Recall of self-subjective information

Table 2 shows a closer look at the recall of self-subjective information (that is, information related to people's response to auto-questionnaires). These results, shown graphically in Fig. 2, indicate decay for the PANAS-N and STAI S particularly, across experimental paradigms, with respective over-estimation of negative emotions (PANAS-N and.

STAI-S). Interestingly, the recall of positive emotions (PANAS-P) was consistently more accurate than the recall of negative ones (PANAS-P).

A comparison of these results (tabulated in the supplementary material) showed significant differences between the Control and VR paradigms for the recall of PANAS-P (U = 195, p < 0.001) and STAI-S (U = 262, p < 0.01). A further difference was detected

Table 2
Recall scores (% and normalised) for Self-Subjective recall.

Paradigm	Parameter (Mean score)	PANAS-P	PANAS-N	STAI-S
Physical	%	13%	34%	−63%
	Normalised	0.0797	0.198	−0.551
Control	%	17%	29%	−64%
	Normalised	0.1143	0.134	−0.4463
VR	%	0%	21%	−65%
	Normalised	−0.039	0.127	−0.6247

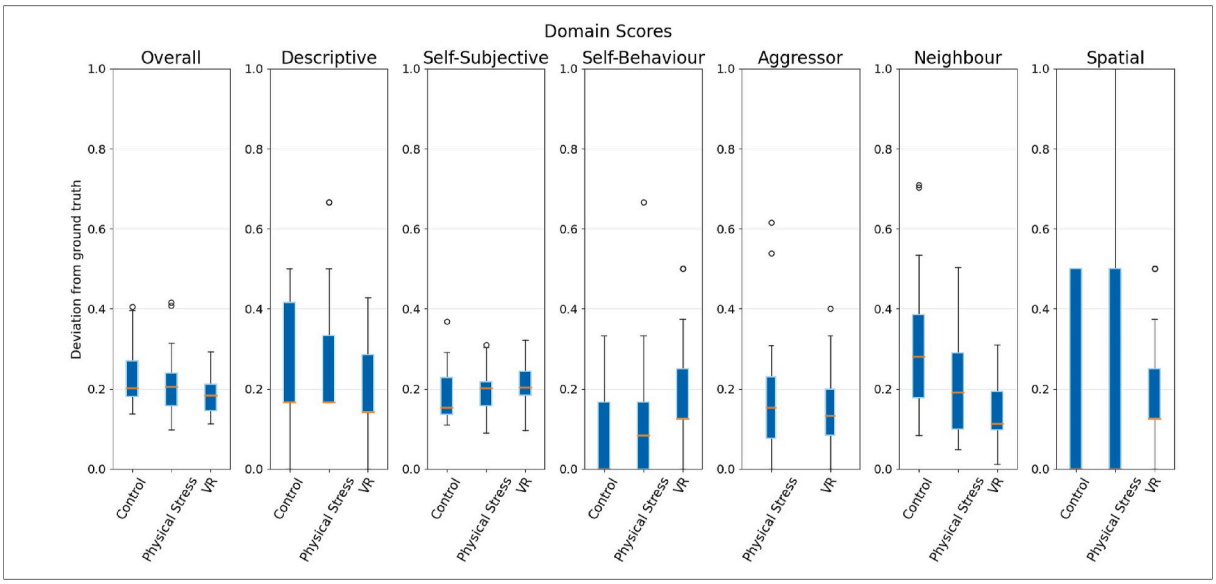


Fig. 1. Domain scores across all paradigms.

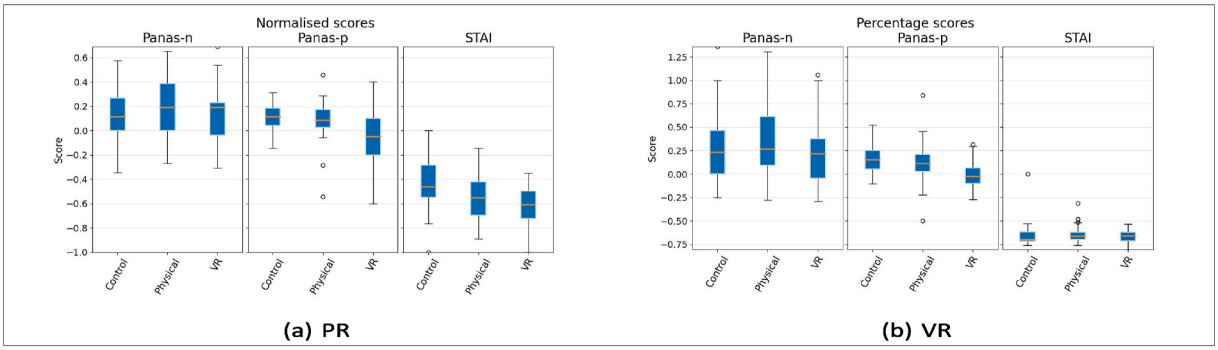


Fig. 2. Recall scores (% and normalised) for Self-Subjective recall.

between the Physical and VR paradigms for PANAS-P ($U = 474$, $p < 0.001$). No statistically significant differences were observed between the Physical and Control paradigms.

3.3. Regression models

To understand the relative importance of the different predictors of recall quality, a full regression model was developed for each domain. The data from Physical and VR groups was combined to make these models, however a further 'VR' predictor variable was used to indicate any statistical differences between the two paradigms. The results of the model are detailed in Table 3, showing the significance of each of the coefficients, as well as the model quality measurements (adjusted r^2, F-statistic and omnibus residual normality tests).

The models suggest a disparate relationship between the predictor variables and the recall quality, with Exercise being the most commonly significant variable, showing an improvement (i.e., a coefficient < 0) in participants' recall of Self-Behaviour ($p < 0.05$), Aggressor ($p < 0.05$), and Neighbour ($p < 0.01$) domains. There is also a significant ($p < 0.05$) relationship between Maximum Cortisol Difference and Aggressor-recall, with higher cortisol differentials resulting in an improved recall quality. Finally, VR participants had a significantly ($p < 0.05$) better recall quality of Neighbours, corresponding to the raw differences detailed above in Table 2.

From Table 3 it can be seen that the models have an overall low level of explanatory power (as the adjusted r^2 values range from 0.0069 to 0.1436) and tend not to be statistically significant (only the Neighbour-focused domain has a significant F-statistic). Interestingly, the residuals seem to be strongly non-normal for three of the domains (Descriptive, Self-Behaviour and Aggressor), as well as for the Spatial domain.

3.4. Confabulation

Table 4 shows the level of confabulation seen across the experimental groups. No statistically significant differences were in the level of confabulation between the different paradigms, or between the combined Physical and Control paradigms.

4. Discussion

It was observed that stress can have a varying effect on an individual's recall quality, depending on the subject domain. When comparing the impact of stress between Control and both VR and Physical paradigms, this study identified a single domain where stressful stimuli improved recall quality (Neighbour), and a further domain where these stimuli appeared to have no effect (Spatial). All other domains showed potentially mixed effects, with differences between VR-Control and PR-Control comparisons. The differences in Neighbour-focused recall could be a result of an increased importance placed on social interactions during emergencies, and would

Table 3
Regression models for specific recall domains.

Coefficient	Overall	Descriptive	Selfsubjective	Selfbehaviour	Aggressor	Neighbour	Spatial
Constant	0.2147**	0.2004	0.1899**	0.3568*	0.3199*	0.2145	0.1794
VR experiment	−0.0308	−0.0705	0.0152	0.0491	−0.0202	−0.0667*	−0.1280
Delta-STAI	8.41e-05	0.0012	−0.0007	−0.0011	0.0002	−0.0004	0.0039
PANAS-N	−0.0002	−0.0014	0.0014	−0.0018	0.0001	0.0009	−0.0032
PANAS-P	0.0020	0.0024	−0.0003	−0.0005	0.0021	0.0023	0.0076
Cortisol	−0.0038	−0.0060	0.0018	−0.0080	−0.0082*	−0.0047	−0.0077
Age	−0.0010	−0.0004	−0.0015	−0.0024	−0.0049	0.0008	−0.0027
Exercise	−0.0089	0.0033	0.0064	−0.0286*	−0.0265*	−0.0287**	−0.0021
Gender (F)	0.0158	0.0460	−0.0108	−0.0028	0.0364	0.0065	0.0177
Adjusted r^2	0.0464	0.0267	0.0161	0.0428	0.0571	0.1436	0.0069
F Statistic	1.4437	1.2641	1.1496	1.4309	1.5834	2.6145*	1.0674
Omnibus test	5.4742	10.6483**	2.1494	16.6080***	14.1794***	3.6719	6.9113*

†$p < 0.1$, *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Table 4
Confabulation proportions.

Paradigm	Confabulation proportion
Physical	0.153
Control	0.09
VR	0.167

suggest that testimonies provided by people following emergency situations are relatively reliable (as compared to when there is no stress involved), insofar as the 'social' dimension of things is concerned.

Interestingly, there are significant overlaps between participant recall within Physical and VR paradigms. There was no significant difference observed between the recall of these paradigms within the following domains: Self-subjective, Aggressor, Spatial. Furthermore, there were only marginally significant differences in the following domains. Overall, Self-behaviour. However, given the number of tests performed, it is possible that this is an artefact of the study, and not a real effect.

This shows that there are multiple areas in which an individual's recollection of their experience in VR would be identical to that of a Physical experiment. There was a strongly significant ($p < 0.001$) improvement in VR recall within the Descriptive domain and a significant ($p < 0.05$) difference in the Neighbour-focused. It is possible that this difference is a product of the VR environment being a simplistic representation of the Physical environment.

The emotional state and demographic (PANAS, delta-STAI and max cortisol difference) of the participants were used as predictors for their recall scores. However, this study did not find any significant relationship between self reported emotional state and quantitative recall accuracy. This is considered highly interesting, as previous research and anecdotal data would suggest that recall quality is improved in some areas and reduced in others. A possible explanation of this is the significant relationship observed based on maximum cortisol difference, providing some evidence to suggest that an increase in stress causes a change in recall quality. The authors suggest that this result is worth further investigation. The only relationships observed between demographic and recall quality revolved around the exercise level of the participants.

The deterioration in Self-Subjective recall is only present when comparing the signed responses of VR and Control experiments. Additionally, during the investigation of the Self-Subjective domain, this study observed several specific effects. When considering the PANAS recall scores, participants tended to overestimate both positive and negative emotional states. However, there was also wide variation in the level of this overestimation. Participants consistently underestimated their STAI-S responses, but there was less variation within these recall scores than with the PANAS, with participant scores lying within approximately 50–70% underestimation change of their original stated scores, while the PANAS scores varied by up to 50% underestimation and 120% overestimation. This study also observed a strongly significant effect where in participants' recall of emotional state, with PANAS-N being overestimated significantly ($p < 0.001$) more than PANAS-P across all paradigms.

When investigating subjective memories specifically, this study has seen that, on average, while the PANAS is on average more reliable through time than the STAI, the STAI deteriorates more consistently across paradigms. Therefore the STAI would be more appropriate for assessing the emotional state of participants through recall after approximately three months.

Finally, the confabulation scores were equivalent across paradigms, suggesting that the degree of memory creation does not depend on the level of stress. However, it is suggested that this may not be the case with more extreme stress levels.

There are limitations to how we can interpret the results shown above. The primary limitation is the number and size of the datasets. There were financial and logistical limitations to the implementation of the experiments, restricting the number of participants, and while best efforts were made to ensure participation in the post-experimental surveys, there were still participants who did not respond.

Another consideration with respect to the size of the datasets is the potential for group-based effects owing to the 10 discrete physical experiments. There was significant preparatory work done in consideration of group membership of the physical experiment participants. This includes randomly assigning participants to the different experimental groups, and assessing their level of group identification through an adapted survey from Ref. [42]. No group identification effects were observed. Furthermore, the VR participants took part in experiments individually, and were exposed to identical stimuli, reducing any group-based differences. One way to analyse this concern would be to perform a leave one-out analysis, and see how robust the results are over the inclusion/exclusion of each group. Unfortunately there is a direct trade-off in statistical explanatory power when investigating these effects against the broader relationships. Given these limitations, and the fact that participants were randomly assigned to different groups in the first instance, this analysis was not performed here. However, it would be as fruitful avenue of research for future studies.

A final obvious limitation in this experimental approach is the difference in the intensity between the proposed situation and any actual threat to life. This was extensively considered in the design phase and again during the ethics approval procedure, concluding that higher stress scenarios are possible experimentally, but the participants would have to be informed of these ahead of the experiment, which would introduce far more biases in the dataset. Instead, the participants were exposed to an unexpected stressor, which quickly became clear that there was no actual threat of harm. It was considered that this was the most appropriate way of studying recall quality from emergencies.

5. Conclusion

This study has investigated the quantitative accuracy of recall of a stressful event, by generating a dataset of recall from which a perfect, 'ground truth' is available. The study then proposed an analysis of this dataset, splitting the responses into different domains. This is a highly pertinent topic for emergency-related study, as most research in these areas involve some form of self-report data collection. The results reported here are therefore relevant for a variety of different scenarios including natural disasters, crowd crushes, and terrorist attacks.

The highlighted results of this study indicate minor differences in recall quality between VR and Physical experiments, suggesting that the subjective experience between the two is similar. Further highlighted results indicate the lack of a self-reported stress-based effect on the quality of participant recall. This result is unusual, as accepted opinion would suggest that recall quality is dependent on the level of stress experienced by the individual. This independence from stress is also shared in the level of confabulation by the participants.

This study has focused on comparing the quantitative accuracy of different recall domains (e.g. recall of an aggressor versus recall of a neighbour). However, there are numerous questions remaining within this approach, including fully assessing the relative accuracies of the different domains, or indeed if other domains are more appropriate for investigation. Future work could consider different domains, or a different paradigm completely. The results of this study should therefore be considered in light of this, and the authors suggest that this research is a step in a larger area of research.

Author contributions statement

A.S. conceived and implemented the initial experiments. A.S. and G.D. developed the survey and performed the analysis. All authors reviewed the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

A.S. was funded by EPSRC grant EP/L016826/1. G.D. was funded by the British Academy (Newton International Fellowship NF171514.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijdrr.2022.103502.

N	Survey items (PR)	Domains	Survey items (VR)	Domains
1	What happened in this experiment?	2	What happened in this experiment?	2
2	How many participants took part in the experiment (including yourself)?	6	How many participants took part in the experiment (including yourself)?	6
3	How many participants were female (including yourself, if appropriate)?	6	How many participants were female (including yourself, if appropriate)?	6
4	What were the instructions given by the aggressor?	2,5	What were the instructions given by the aggressor?	2,5
5	Where were you standing when you noticed the aggressor?	2,5	What was the colour of the aggressor's shirt?	2,5
6	Which gender was the aggressor?	2,5	What was the colour of the aggressor's hair?	2,5
7	What was the colour of the aggressor's hair?	2,5	Which gender was the aggressor?	2,5
8	What was used to mark out the physical boundary between the two parts of the gym (in blue)?	2,8	Where were you standing during the first (nonstressful) part of the experiment?	2,8
9	What was your very first (e.g. first 5 s) behaviour when you noticed the aggressor?	4,5,8	Where were you standing when you noticed the aggressor?	2,5,8
10	I *did/did not* verbally confront the aggressor	4,5	I *did/did not* verbally confront the aggressor	4,5
11	I *did/did not* try dodging the aggressor when they started pursuing me	4,5	I *did/did not* try dodging the aggressor when they started pursuing me	4,5
12	I *did/did not* help a participant that the aggressor had approached	4,5,6	I *did/did not* help a participant that the aggressor had approached	4,5,6
13	I *did/did not* take another participant by the hand	4,6	I *did/did not* take another participant by the hand	4,6
14	I *did/did not* confront the aggressor physically	4,5	I *did/did not* confront the aggressor physically	4,5
15			I *did/did not* try to leave the room	4,8
16			I *did/did not* go around the central column	4,8

(*continued on next page*)

N	Survey items (PR)	Domains	Survey items (VR)	Domains
17			I *did/did not* go behind the dividing wall	4,8
18			I *did/did not* go behind the central column	4,8
19	*No one/at least one person * verbally confronted the aggressor	5,6	*No one/at least one person * verbally confronted the aggressor	5,6
20	*No one/at least one person * tried dodging the aggressor when they started pursuing him/her	5,6	*No one/at least one person * tried dodging the aggressor when they started pursuing him/her	5,6
21	*No one/at least one pair of other participants* were holding hands	6	*No one/at least one pair of other participants* were holding hands	6
22	*No one/at least one person * helped a participant that the aggressor had approached	5,6	*No one/at least one person * helped a participant that the aggressor had approached	5,6
23	*No one/at least one person * confronted the aggressor physically	5,6	*No one/at least one person * confronted the aggressor physically	5,6
24			*Nobody/at least one other participant* went behind the dividing wall	6,8
25			*Nobody/at least one other participant* went behind the central column	5,8
26			*No one/at least one other participant* tried to leave the room	5,8
27	Group Identity	3	Group Identity	3
28	STAI	3	STAI	3
29	PANAS	3	PANAS	3
30			How much did you believe the avatars were real people?	3

Average absolute normalised score per domain

Paradigm	Metric	Overall	Descriptive	Self-subjective	Self-behaviour	Aggressor	Neighbour	Spatial
Control	Mean	0.236	0.281	0.188	0.096	N/A	0.305	0.184
	SD	0.078	0.158	0.072	0.128	N/A	0.186	0.248
	Min	0.138	0.000	0.110	0.000	N/A	0.083	0.000
	Max	0.405	0.500	0.368	0.333	N/A	0.710	0.500
PR Stress	Mean	0.205	0.269	0.184	0.111	0.162	0.207	0.278
	SD	0.075	0.151	0.075	0.144	0.141	0.124	0.326
	Min	0.066	0.167	0.000	0.000	0.000	0.049	0.000
	Max	0.415	0.667	0.309	0.667	0.615	0.503	1.000
VR	Mean	0.182	0.189	0.207	0.160	0.155	0.140	0.182
	SD	0.047	0.117	0.061	0.125	0.080	0.077	0.141
	Min	0.108	0.000	0.000	0.000	0.000	0.011	0.000
	Max	0.293	0.429	0.322	0.500	0.400	0.309	0.500

Mann Whitney test results on domain scores

Paradigm	Metric	Overall	Descriptive	Self-subjective	Self-behaviour	Aggressor	Neighbour	Spatial
PR-VR	Statistic	665	370	709.5	668	771	587	809.5
	p-value	0.064†	5.51e-06***	0.135	0.062†	0.297	0.012*	0.431
Control-VR	Statistic	256	198.5	323	326	N/A	172	365.5
	p-value	0.005**	0.00018***	0.051†	0.0495*	N/A	6.81e-05***	0.145
Control-PR	Statistic	274	321	309	324.5	N/A	226.5	296
	p-value	0.116	0.343	0.283	0.369	N/A	0.021*	0.179

Differences in raw scores of subjective recall

Paradigms		PANAS-P	PANAS-N	STAI-S
Physical-VR	Statistic	474.0	720.5	685.5
	p-value	0.0007***	0.198	0.119
Control-VR	Statistic	195.0	419.5	262.0
	p-value	0.0003***	0.456	0.008**
Physical-Control	Statistic	280.0	313.0	282.0
	p-value	0.137	0.306	0.146

References

[1] Global terrorism database. Accessed on 9/November/2017. [Online]. Available: https://www.start.umd.edu/gtd/.
[2] P. Nesser, A. Stenersen, The modus operandi of jihadi terrorists in europe, Perspectives on Terrorism 8 (2015) 1.
[3] J. Drury, The role of social identity processes in mass emergency behaviour: an integrative review, Eur. Rev. Soc. Psychol. 29 (1) (2018) 38–81 https://doi.org/10.1080/10463283.2018.1471948 [Online]. Available:
[4] C. Arteaga, J. Park, Building design and its effect on evacuation efficiency and casualty levels during an indoor active shooter incident, Saf. Sci. 127 (2020)

104692 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925753520300898.
[5] A.R. Mawson, Understanding mass panic and other collective responses to threat and disaster, Psychiatr. Interpers. Biol. Process. 68 (2) (2005) 95–113.
[6] M. Kobes, I. Helsloot, B. de Vries, J.G. Post, N. Oberijé, K. Groenewegen, Way finding during fire evacuation; an analysis of unannounced fire drills in a hotel at night, Build. Environ. 45 (3) (2010) 537–548 [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0360132309001759.
[7] G. Dezecache, J.R. Martin, C. Tessier, L. Safra, V. Pitron, P. Nuss, J. Grèzes, Nature and determinants of social actions during a mass shooting, PLoS One 16 (12) (2021) e0260392.
[8] G. Dezecache, Human collective reactions to threat, Wiley Interdiscip Rev Cogn Sci 6 (3) (2015) 209–219.
[9] D. Schweingruber, R.T. Wohlstein, "The madding crowd goes to school: myths about crowds in introductory sociology textbooks, Teach. Sociol. 33 (2) (2005) 136–153 [Online]. Available: http://www.jstor.org/stable/4127520.
[10] A. Shipman, A. Majumdar, Fear in humans: a glimpse into the crowd-modeling perspective, Transport. Res. Rec. 2672 (1) (2018) 183–197 https://doi.org/10.1177/0361198118787343 [Online]. Available:
[11] A. Bartolucci, C. Casareale, J. Drury, Cooperative and Competitive Behaviour Among Passengers during the Costa Concordia Disaster," Safety Science, 134, 2021 [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925753520304525. 105055.
[12] J. Drury, C. Cocking, S. Reicher, Everyone for themselves? a comparative study of crowd solidarity among emergency survivors, Br. J. Soc. Psychol. 48 (Pt 3) (2009) 487–506 [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/18789185.
[13] J. Drury, C. Cocking, S. Reicher, A. Burton, D. Schofield, A. Hardwick, D. Graham, P. Langston, "Cooperation versus competition in a mass emergency evacuation: a new laboratory simulation and a new theoretical model, Behav. Res. Methods 41 (3) (2009) [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/19587213. 957–70.
[14] M. Levine, P.J. Taylor, R. Best, Third parties, violence, and conflict resolution: the role of group size and collective action in the microregulation of violence, Psychol. Sci. 22 (3) (2011) 406–412 https://doi.org/10.1177/0956797611398495, pMID: 21303991. [Online]. Available:
[15] L.S. Liebst, R. Philpot, M. Levine, M.R. Lindegaard, Cross-national Cctv Footage Shows Low Victimization Risk for Bystander Interveners in Public Conflicts, vol. 11, 2021, pp. 11–18.
[16] R. Philpot, L.S. Liebst, M. Levine, W. Bernasco, M.R. Lindegaard, Would I Be Helped? Cross-National Cctv Footage Shows that Intervention Is the Norm in Public Conflicts, vol. 75, 2020, pp. 66–75.
[17] S.J. Sharman, M.B. Powell, A comparison of adult witnesses' suggestibility across various types of leading questions, Appl. Cognit. Psychol. 26 (1) (2012) 48–53.
[18] T. Van de Mortel, Faking it: social desirability response bias in self-report research, Aust. J. Adv. Nurs. 25 (2008) 40–48.
[19] P.M. Rodriguez Mosquera, A.K. Uskul, S.E. Cross, The centrality of social image in social psychology, Eur. J. Soc. Psychol. 41 (4) (2011) 403–410 https://doi.org/10.1002/ejsp.820 [Online]. Available:
[20] P. Verger, M. Rotily, C. Hunault, J. Brenot, E. Baruffol, D. Bard, Assessment of exposure to a flood disaster in a mental-health study, J. Expo. Anal. Environ. Epidemiol. 13 (2003) 436–442.
[21] B. Hunter, M. Gray, B. Edwards, The use of social surveys to measure drought and the impact of drought, Soc. Indicat. Res. 113 (8) (2013).
[22] B. Edwards, M. Gray, J.B. Borja, Measuring natural hazard-related disasters through self-reports, International Journal of Disaster Risk Science 12 (4) (Jun. 2021) 540–552 https://doi.org/10.1007/s13753-021-00359-1 [Online]. Available:
[23] C.W. Quaedflieg, L. Schwabe, Memory dynamics under stress, Memory 26 (3) (2018) 364–376.
[24] J. Payne, L. Nadel, J. Allen, K. Thomas, W. Jacobs, The effects of experimentally induced stress on false recognition, Memory 10 (2002) 1–6.
[25] L. Schwabe, O. Wolf, Stress prompts habit behavior in humans, J. Neurosci. : the official journal of the Society for Neuroscience 29 (2009) 7191–7198.
[26] K. Deffenbacher, B. Bornstein, S. Penrod, E. McGorty, A meta-analytic review of the effects of high stress on eyewitness memory, Law Hum. Behav. 28 (2005) 687–706.
[27] E.F. Loftus, The malleability of human memory: information introduced after we view an incident can transform memory, Am. Sci. 67 (3) (1979) 312–320 [Online]. Available: http://www.jstor.org/stable/27849223.
[28] E. Loftus, Eyewitness Testimony, Applied Cognitive Psychology, 2019.
[29] S. Christianson, Emotional stress and eyewitness memory: a critical review, Psychol. Bull. 112 (2) (1992) 284–309.
[30] S. Christianson, B. Hübinette, Hands up a study of witnesses' emotional reactions and memories associated with bank robberies, Appl. Cognit. Psychol. 7 (5) (1993) 365–379.
[31] T. Smeets, T. Giesbrecht, M. Jelicic, H. Merckelbach, Context-dependent enhancement of declarative memory performance following acute psychosocial stress, Biol. Psychol. 76 (2007) 116–123.
[32] C. Grady, S. Sarraf, C. Saverino, K. Campbell, Age differences in the functional interactions among the default, frontoparietal control, and dorsal attention networks, Neurobiol. Aging 41 (2016) 159–172.
[33] O. Hardt, K. Nader, L. Nadel, Decay happens: the role of active forgetting in memory, Trends Cognit. Sci. 17 (3) (2013) 111–120.
[34] S.M. Hoscheidt, K.S. LaBar, L. Ryan, W.J. Jacobs, L. Nadel, Encoding negative events under stress: high subjective arousal is related to accurate emotional memory despite misinformation exposure, Neurobiol. Learn. Mem. 112 (2014) 237–247 stress and the regulation of memory: From basic mechanisms to clinical implications. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1074742713001858.
[35] C. Laney, E.F. Loftus, Traumatic memories are not necessarily accurate memories, Can. J. Psychiatr. 50 (13) (2005) 823–828 https://doi.org/10.1177/070674370505001303, pMID: 16483115. [Online]. Available:
[36] D. Strange, M.K.T. Takarangi, Memory distortion for traumatic events: the role of mental imagery, Front. Psychiatr. 6 (2015) [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyt.2015.00027.
[37] A. Ehlers, D.M. Clark, A cognitive model of posttraumatic stress disorder, Behav. Res. Ther. 38 (4) (2000) 319–345.
[38] A. Shipman, A generalised methodology for the investigation of human behavioural responses to hostile attacks. Thesis, 2021 Imperial College London. [Online]. Available: http://hdl.handle.net/10044/1/91981.
[39] Pr survey. Accessed on 04/02/2022. [Online]. Available: https://imperial.eu.qualtrics.com/jfe/form/SV_4GZTAQvkLmbkfPv.
[40] Vr survey. Accessed on 04/02/2022. [Online]. Available: https://imperial.eu.qualtrics.com/jfe/form/SV_8dm0pgOLh6cbR6R.
[41] G. Dezecache, J. Grèzes, C.D. Dahl, The nature and distribution of affiliative behaviour during exposure to mild threat, R. Soc. Open Sci. 4 (8) (2017) https://doi.org/10.1098/rsos.170265 [Online]. Available: 170265, Aug.
[42] W. Leach Colin, M.R. Mosquera Patricia, L.W. Vliek Michael, E. Hirt, Group devaluation and group identification, J. Soc. Issues 66 (3) (2010) 535–552 https://doi.org/10.1111/j.1540-4560.2010.01661.x [Online]. Available: