Does US expectancy mediate the additive effects of CS-US pairings on contingency instructions? Results from subjective, psychophysiological and neural measures

Word count: 4862

Gaëtan Mertens[1], Senne Braem[2], Manuel Kuhn[3], Tina B. Lonsdorf[3], Marcel A. van den Hout[1], and Iris M. Engelhard[1]

[1]Department of Clinical Psychology, Utrecht University, Utrecht, the Netherlands

[2]Department of Experimental Psychology, Ghent University, Belgium

[3]Institute for Systems Neuroscience, University Clinic Hamburg-Eppendorf, Germany

Correspondence concerning this article should be addressed to Gaëtan Mertens, Department of Clinical Psychology, Heidelberglaan 1, room H1.29, Utrecht University, 3584CS Utrecht, the Netherlands.

E-mail: g.mertens@uu.nl

Tel: +31 30 253 75 53

Highlights

- CS-US pairings can add to the effects of CS-US contingency instructions.
- We investigate whether US expectancy ratings can account for these effects.
- Mediational analyses did not support this hypothesis.
- Exploratory analyses revealed that fear ratings do mediate these effects.
- We discuss the validity of US expectancy and fear ratings.

Abstract

Verbal instructions are a powerful pathway to learn new fear relations, and an important question has been what fear experience can still add to the effect of such instructions. Therefore, in previous studies, we investigated the effects of pairings between conditioned stimuli (CS) and unconditioned stimuli (US) after CS-US contingency instructions. Although these studies found that CS-US pairings do indeed add to the effects of contingency instructions on subjective, psychophysiological and neural measures of conditioned fear, they also produce increases in US expectancy ratings. In the current report we address whether these enhanced US expectancy ratings can account for the additive effects of CS-US pairings as suggested by expectancy models of fear conditioning. To address this question we made use of pathway models to investigate mediation in within-subjects designs. Our results demonstrate that US *expectancy* ratings do not mediate the effects of CS-US pairings on fear ratings, the startle reflex or amygdala activation pattern similarity. Additional exploratory analyses, however, revealed that subjective *fear* ratings do explain the effects of CS-US pairings on the other measures. We discuss how these results relate to expectancy models of fear conditioning and what they implicate for the validity of US expectancy and fear ratings.

Keywords: Expectancy; Conditioning; Subjective fear; Psychophysiology

Does US expectancy mediate the additive effects of CS-US pairings to contingency instructions? Results from subjective, psychophysiological and neural measures

Humans display the adaptive ability to quickly learn to fear and avoid stimuli that predict possible harmful events. We are capable of learning this contingency not only through the pairing of initially neutral conditioned stimuli (CSs) and aversive unconditioned stimuli (US; i.e., fear conditioning), but also through verbal instructions and social observation. Although the delineation of these different pathways has been described at least 40 years ago (Rachman, 1977), the interaction between the different pathways is still not well understood.

In recent studies we have addressed this interaction between the verbal and experiential pathway (Braem et al., 2017; Mertens, Kuhn, et al., 2016; Mertens, Raes, & De Houwer, 2016; Raes, De Houwer, De Schryver, Brass, & Kalisch, 2014). Specifically, these studies investigated whether CS-US pairings (i.e., conditioning trials) add to the effect of clear and believable verbal contingency instructions. Therefore, participants first went through a conditioning phase in which one CS (CS+P) was paired with a US (a mild electric shock) while another CS (CS+U) was not paired (or: unpaired) with the US. Importantly, participants were told at the outset of the experiment that the CS+U would not be followed by the US in the first phase, but would be followed by the US in the second test phase, and participants were reminded of these instructions in between phases. In reality, however, none of the CSs were followed by the US in the test phase, which ensured that the conditioned response for the CS+U was purely based on instructions. Across all four studies, we found clear evidence that the CS+P elicited slightly larger fear responses than the CS+U (i.e., subjective fear ratings, potentiated startle response and amygdala activation pattern similarity; but not skin conductance responses), suggesting that CS-

US pairings add to the effect of verbal instructions (Braem et al., 2017; Mertens, Kuhn, et al., 2016; Mertens, Raes, et al., 2016; Raes et al., 2014).

Our studies also showed that CS-US pairings did not only influence fear responses, but also increased participants' expectancy ratings for the CS+P. An unaddressed question in our previous studies is whether these increased expectancies could account for the increased fear responses. This would be expected on the basis of several important theories of fear conditioning. Specifically, according to Davey's expectancy model of fear conditioning (Davey, 1992), conditioned fear responses reflect participants' expectancy and evaluation of the US. Similar models have been proposed by Reiss (1980), Lovibond (2011) and Dawson and Furedy (1976). Hence, according to these models one may predict that the increased expectancy ratings due to CS-US pairings mediate the effects of CS-US pairings on the fear responses. Alternatively, other models of fear conditioning have argued that CS-US pairings can install memory associations that are independent of language and expectancies (LeDoux, 2014; Öhman & Mineka, 2001; Olsson & Phelps, 2007). According to these latter theories, CS-US pairings may increase fearful responses without necessarily altering expectancy ratings (e.g., Mineka & Öhman, 2002).

In order to address these competing predictions regarding the mediating role of US expectancies for explaining the additive effects of CS-US pairings, we have re-analyzed data from our prior studies using recent methods for performing mediation analyses for within-subjects designs (see Montoya & Hayes, 2017). If the additive effects of CS-US pairings on fear measures (fear ratings, fear potentiated startle and amygdala activation pattern similarity) are explained by increases in US expectancy ratings, mediation analysis should indicate that US expectancy ratings significantly mediate the effects of CS-US pairings on these measures.

Alternatively, if US expectancy ratings do not explain the additive effects of CS-US pairings on fear measures, the effects of CS-US pairings on fear measures should remain present even when partialling out the variance related to US expectancy ratings.

Method

Participants

To address the hypotheses stated above, we have re-analyzed the data of the four previous studies that have investigated the additive effects of CS-US pairings to contingency instructions (Braem et al., 2017; Mertens, Kuhn, et al., 2016; Mertens, Raes, et al., 2016; Raes et al., 2014). These samples consisted of healthy university students (Braem et al.: N = 20; Mertens, Kuhn, et al.: N = 36; Mertens, Raes, et al.: N = 36; Raes et al., N = 31[1]).

Materials and procedure

The procedure of these studies has been extensively described in the original studies. In brief, participants took part in a single session fear conditioning experiment. In a first phase, participants were informed about the contingency between pictures of snow fractals (or pictures of fear-relevant and fear-irrelevant animals, see Mertens, Raes, et al., 2016, Experiment 2) and an electric stimulation via instructions on the computer screen. They were told that two of these snow fractals would sometimes be followed by an electric stimulation during the experiment, whereas a third fractal would never be followed by the stimulation. Furthermore, participants were informed that in the first part of the experiment, some of the electric stimulations would be

[1] One participant from the Raes et al. (2014) dataset was excluded after inspection of the data due to her ratings being an extreme outlier. That is, this participant indicated a difference in US expectancy between CS+P and CS+U of 6 (on a 9-point scale) and a difference in fear ratings between CS+P and CS+U of -6 (also on a 9-point scale). Excluding this participant from the data changed the correlation between US expectancy ratings and fear ratings from .113 to .638 (see the Supplementary Material).

replaced by a picture of a lightning bolt in order not to expose them to too many electric stimulation. During this first phase, the snow fractals were presented on the computer screen for 8 seconds. One of the fractals was sometimes (i.e., on 33% of the trials) followed at offset by an electric stimulation (CS+P), whereas another fractal was sometimes (also on 33% of the trials) followed by a picture of a lightning bolt (CS+U). A third fractal was never paired with the stimulation or with the picture of the lightning bolt (CS-).

Following this first phase, participants were told that in the next phase no more replacements would be presented (i.e., the picture of a lightning bolt), and that the two fractals (referred to in the instructions in the first phase) would now actually be followed by the electrical stimulation. They were further informed that the third fractal would still not be followed by the stimulation. The procedure of this second phase (i.e., the crucial test phase) was identical to the previous phase with the exception that the electrical stimulation and the picture of the lightning bolt were never presented.

Each phase was interrupted three times by a ratings block in which participants had to rate their subjective fear levels (“How much fear did you experience while looking at this figure?”) and US expectancy (“To what extent did you expect an electro-tactile stimulation while seeing this figure?”) for the three different snow fractals on 9-point Likert scales (as further explained in Braem et al., 2017, not all ratings were assessed in the first subjects of that study, resulting in a slightly smaller sample for some of the analyses below). Besides these subjective ratings, we have also collected skin conductance responses (SCRs; Mertens, Kuhn, et al., 2016; Mertens, Raes, et al., 2016; Raes et al., 2014), potentiation of the startle reflex (Mertens, Kuhn, et al., 2016) and the fMRI BOLD signal (Braem et al., 2017) during the fractal presentations.

The crucial comparison was during the test phase, between the fractal that had been paired with the stimulation (CS+P) and the fractal that was only paired with the picture of a lightning bolt (CS+U). More specifically, the analyses zoomed in on fear responses during the first three trials of the test phase, given that we expect the effects of prior CS-US pairings to be most pronounced during these first few trials because they are less affected by extinction due to non-reinforcement of the CSs during the test phase. Furthermore, also the believability of the contingency instructions (i.e., that CS+U will now also be followed by electrical stimulations) is unlikely affected by the non-reinforcement of the CSs during these first three trials of the test phase.

Data preprocessing and analysis

Preprocessing of the fear responses. Scoring of the psychophysiological responses has been extensively described in the previous reports. In brief, startle responses (or: fear potentiated startle, FPS) were scored by taking the maximum amplitude in the 20-120 milliseconds time window after the startle probe onset (Mertens, Kuhn, et al., 2016). Amygdala activation pattern similarity was calculated as the similarity in voxel pattern activation between CS+P presentation during the training phase, and CS+P or CS+U presentation during the testing phase (Multi-Voxel Pattern Analysis; see Braem et al., 2017 for an extensive description of this approach). Finally, skin conductance was also measured but will not be considered here because no effects of CS-US pairings were found for this measure in any of our studies (Mertens, Kuhn, et al., 2016; Mertens, Raes, et al., 2016; Raes et al., 2014).

Statistical analyses. To investigate mediation of the additive effects of CS-US pairings by US expectancy ratings, we have performed mediational analyses using the MEMORE syntax developed by Montoya and Hayes (2017) in SPSS (version 24.0). This code provides the

pathway coefficients, the standard error and the 95% confidence interval for the direct (i.e., the effect of a factor when controlled for the shared variance with the mediator) and the indirect (i.e., the mediation effect) pathways in a mediation model. Mediation is established in the path analytic framework when the confidence interval for the pathway coefficient of the indirect pathway does not include zero (Montoya & Hayes, 2017; Preacher & Hayes, 2004). Particularly, we investigated whether the effect of CS type (CS+P versus CS+U) during the first block of the test phase on the different fear measures (i.e., the direct effect of CS type) could be accounted by US expectancy ratings (i.e., the indirect effect of CS type through US expectancy). Pathways were estimated using 10,000 bootstrap samples. A table containing the correlations between the different variables, a table containing the intraclass correlations of the different measures throughout the test phase (providing an index of the reliability of these measures; e.g., Shechner et al., 2015), and a table containing the model fits of the different models (estimated using MPlus 8.0) are included in the Supplementary Materials.

Results

Main analysis

The pathway coefficients, the standard errors, and the 95% confidence intervals of the mediation analyses of the effects of CS-US pairings on the different measures of fear with US expectancy ratings as a mediator are summarized in Table 1. Two important findings can be highlighted from this table. First, the direct pathway (i.e., the effects of CS-US pairings while controlling for US expectancy ratings) was significant for fear ratings in the Raes et al. (2014) and Mertens, Kuhn, et al. (2016) datasets, and marginally significant in the Mertens, Raes, et al. (2016) dataset. Similarly, the direct pathway was significant for FPS in the Mertens, Kuhn, et al.

(2016) dataset. These results indicate that US expectancy ratings did not fully account for the effects of CS-US pairings on these measures. Second, mediation of the effect of CS-US pairings on fear ratings by US expectancy ratings was established in all datasets (though only marginally significantly so in the Braem et al., 2017, dataset; see the results of the indirect pathway). However, this mediation was only partial because the direct pathways remained (marginally) significant in these datasets when controlling for US expectancy (see above and Table 1, though again except for the Braem et al. dataset). Furthermore, US expectancy did not mediate the effects of CS-US pairings on either FPS or amygdala activation pattern similarity.

Table 1. Results of the mediation analyses of the effects of CS-US pairings on fear ratings, fear

potentiated startle and right amygdala activation similarity with US expectancy ratings as the mediator. Pathway coefficients, SEs and 95% confidence intervals of the direct (i.e., the effects of CS-US pairings controlled for US expectancy ratings) and indirect pathway (i.e., mediation of the effects of CS-US pairings by US expectancy ratings) are reported.

Dependent variable	Direct pathway	Indirect pathway
Raes et al. (2014)		
Fear ratings	B = 0.514 SE = 0.242 CI = [0.018, 1.010]*	B = 0.422 SE = 0.272 CI = [0.027, 1.062]*
Mertens, Kuhn, et al. (2016)		
Fear ratings		
	B = 0.386 SE = 0.146	B = 0.475 SE = 0.171
FPS	CI = [0.088, 0.684]*	CI = [0.173, 0.839]*
	B = 3.644 SE = 1.693	B = -0.036 SE = 0.725
	CI = [0.195, 7.092]*	CI = [-1.250, 1.666]
Mertens, Raes, et al. (2016)		
Fear-irrelevant CSs		
Fear ratings		
	B = 0.525 SE = 0.283	B = 1.142 SE = 0.338
	CI = [-0.051, 1.101]$^{+}$	CI = [0.534, 1.839]*
Fear-relevant CSs		
Fear ratings		
	B = 0.570 SE = 0.313	B = 0.875 SE = 0.434
	CI = [-0.067, 1.206]$^{+}$	CI = [0.198, 1.848]*
Braem et al. (2017)		
Fear ratings		
	B = 0.053 SE = 0.080	B = 0.332 SE = 0.206
Inter-CS activation similarity	CI = [-0.117, 0.222]	CI = [-0.012, 0.790]$^{+}$
right amygdala	B = 0.129 SE = 0.097	B = 0.076 SE = 0.058
	CI = [-0.077, 0.334]	CI = [-0.030, 0.198]

Note: * indicates that zero fell outside the 95% CI; $^{+}$ indicates that zero fell outside the 90% CI.

Exploratory analyses

To further explore the relationships between the different dependent variables, we have conducted additional mediation analyses. First, it may be argued that fear ratings more specifically capture subjectively experienced fear compared to US expectancy ratings. Particularly, US expectancy ratings are focused on the expectation of the US, but not the evaluation of the US. Fear, however, is likely a function of both the expectation and evaluation of the US (e.g., Davey, 1992), which may be better captured by fear ratings. Hence, fear ratings may better capture the subjectively experienced fear of participants and may therefore be a more appropriate mediator for the effects of CS-US pairings on the other measures of fear. Second, according to some models (e.g., LeDoux, 2014; Mineka & Öhman, 2002), changes in psychophysiological fear responses may cause changes in US expectancy ratings and fear ratings. In this case, psychophysiological responses (startle, amygdala activation pattern similarity) should mediate the effects of CS-US pairings on US expectancy and fear ratings, rather than the other way around. To evaluate these predictions, we have run additional mediation models with fear ratings, fear potentiated startle, and amygdala activation pattern similarity as mediators and the other measures as outcomes. The results of these analyses are summarized in Tables 2 to 4.

Fear ratings. In contrast to the results from US expectancy ratings, the direct pathway of the effects of CS-US pairings on the different measures of fear was not significant for any of the measures or datasets with fear ratings included as a mediator in the models (see Table 2). This result thus indicates that the effects of CS-US pairings was no longer observed for any of the measures when controlling for the shared variance with fear ratings. Furthermore, fear ratings completely mediated the effect of CS-US pairings on US expectancy ratings in all datasets, and

marginally significantly mediated the effect of CS-US pairings on right amygdala activation pattern similarity (see Table 2).

Table 2. Results of the mediation analyses of the effects of CS-US pairings on US expectancy ratings, fear potentiated startle and right amygdala activation similarity with fear ratings as the mediator. Pathway coefficients, SEs and 95% confidence intervals of the direct (i.e., the effects of CS-US pairings controlled for fear ratings) and indirect pathway (i.e., mediation of the effects of CS-US pairings by fear ratings) are reported.

Dependent variable	Direct pathway	Indirect pathway
Raes et al. (2014)		
US expectancy ratings	B = 0.071 SE = 0.251 CI = [-0.443, 0.584]	B = 0.575 SE = 0.266 CI = [0.101, 1.145]*
Mertens, Kuhn, et al. (2016)		
US expectancy ratings	B = -0.025 SE = 0.201 CI = [-0.434, 0.384]	B = 0.803 SE = 0.212 CI = [0.434, 1.255]*
FPS	B = 2.265 SE = 1.803 CI = [-1.407, 5.937]	B = 1.343 SE = 1.127 CI = [-1.070, 3.358]
Mertens, Raes, et al. (2016)		
Fear-irrelevant CSs		
US expectancy ratings	B = 0.432 SE = 0.309 CI = [-0.196, 1.060]	B = 1.207 SE = 0.349 CI = [0.505, 1.874]*
Fear-relevant CSs		
US expectancy ratings	B = 0.444 SE = 0.306 CI = [-0.178, 1.068]	B = 1.028 SE = 0.364 CI = [0.322, 1.751]*
Braem et al. (2017)		
US expectancy ratings	B = 0.004 SE = 0.073 CI = [-0.151, 0.158]	B = 0.325 SE = 0.151 CI = [0.032, 0.622]*
Inter-CS activation similarity right amygdala	B = 0.112 SE = 0.092 CI = [-0.082, 0.307]	B = 0.092 SE = 0.054 CI = [-0.003, 0.202]$^{+}$

Note: * indicates that zero fell outside the 95% CI; $^{+}$ indicates that zero fell outside the 90% CI.

Fear potentiated startle. The results from the mediational analysis of the effects of CS-US pairings with FPS as a mediator are summarized in Table 3 (this only concerns the Mertens, Kuhn, et al., 2016, dataset). The results of this analysis indicate that the effects of CS-US pairings on US expectancy and fear ratings remained significant when controlling for FPS (see the results of the direct pathway in Table 3). Furthermore, no evidence was obtained for a mediational effect for either US expectancy or fear ratings by FPS (see the results of the indirect pathway in Table 3).

Table 3. Results of the mediation analyses of the effects of CS-US pairings on US expectancy ratings and fear ratings with fear potentiated startle as the mediator. Pathway coefficients, SEs and 95% confidence intervals of the direct (i.e., the effects of CS-US pairings controlled for fear potentiated startle) and indirect pathway (i.e., mediation of the effects of CS-US pairings by fear potentiated startle) are reported.

Dependent variable	Direct pathway	Indirect pathway
Mertens, Kuhn, et al. (2016)		
US expectancy ratings	B = 0.792	B = -0.021
	SE = 0.271	SE = 0.094
	CI = [0.240, 1.344]*	CI = [-0.234, 0.151]
Fear ratings	B = 0.760	B = 0.097
	SE = 0.213	SE = 0.108
	CI = [0.327, 1.193]*	CI = [-0.074, 0.358]

Note: * indicates that zero fell outside the 95% CI.

Amygdala activation pattern similarity. The results from the mediational analysis of the effects of CS-US pairings with amygdala activation pattern similarity as a mediator are summarized in Table 4 (this only concerns the Braem et al., 2017, dataset). The results indicate that the effects of CS-US pairings on US expectancy and fear ratings (i.e., the direct pathway) was not significant when controlled for amygdala activation pattern similarity (note that the

effects of CS-US pairings was marginally significant, $p = .052$, for fear ratings and not significant, $p = .108$, for US expectancy ratings without controlling for amygdala activation pattern similarity). Furthermore, no evidence for a mediation of the effect of CS-US pairings for either US expectancy ratings or fear ratings by amygdala activation pattern similarity was observed (see the indirect pathway in Table 4).

Table 4. Results of the mediation analyses of the effects of CS-US pairings on US expectancy ratings and fear ratings with right amygdala activation similarity as the mediator. Pathway coefficients, SEs and 95% confidence intervals of the direct (i.e., the effects of CS-US pairings controlled for right amygdala activation similarity) and indirect pathway (i.e., mediation of the effects of CS-US pairings by right amygdala activation similarity) are reported.

Dependent variable	Direct pathway	Indirect pathway
Braem et al. (2017)		
US expectancy ratings	B = 0.172	B = 0.157
	SE = 0.194	SE = 0.131
	CI = [-0.237, 0.581]	CI = [-0.059, 0.453]
Fear ratings	B = 0.208	B = 0.176
	SE = 0.205	SE = 0.143
	CI = [-0.224, 0.640]	CI = [-0.035, 0.520]

Discussion

Prior studies from our labs have demonstrated that CS-US pairings can add to the effect of clear contingency instructions. With the analyses reported here we investigated whether enhanced US expectancy ratings due to CS-US pairings could account for the additive effect of the CS-US pairings on fear responses (and vice versa). Our results can be summarized with three main findings: First, US expectancy ratings did not fully account for the additive effect of CS-US pairings on fear responses. That is, after controlling for US expectancy ratings, the additive effects of CS-US pairings on fear ratings and FPS remained significant (see the direct pathway in

Table 1). Furthermore, our results indicate that US expectancy ratings only partially mediated the additive effect of CS-US pairings on fear ratings, and did not mediate this effect for FPS and amygdala activation pattern similarity. Second, analyses in which we included fear ratings as a mediator showed that fear ratings mediated the additive effects of CS-US pairings on US expectancy ratings and amygdala activation pattern similarity (though only marginally so for amygdala activation pattern similarity; see Table 2). Third and final, we did not find evidence in our data that changes in psychophysiological measures of fear (FPS or amygdala activation pattern similarity) mediated additive effects of CS-US pairings on either US expectancy or fear ratings. In the remainder of the discussion, we will relate these results to the different models of fear conditioning presented in the introduction, and discuss the implications of our results for the validity of US expectancy ratings and fear ratings in fear conditioning research.

A first theoretical consideration is that our results seem to contradict expectancy models of fear conditioning (Davey, 1992; Lovibond, 2011; Reiss, 1980). That is, according to these models it would be expected that changes in expectancy ratings mediate the additive effects of CS-US pairings on other measures of conditioned fear. This prediction was not supported by our data. Furthermore, an alternative model in which physiological (rather than subjective) fear responses drive changes in subjectively experienced fear (both as measured with fear ratings and, to a lesser extent, US expectancy ratings) (LeDoux, 2014; Mineka & Öhman, 2002a) was also not supported by our analyses. That is, neither for FPS nor right amygdala activation pattern similarity did we find evidence that they significantly mediated the additive effects of CS-US pairings for either fear ratings or US expectancy ratings. The only model that was consistently found over the different experiments and measures is that fear ratings mediate the additive effects of CS-US pairings for the other measures.

One reason why we did not find evidence for mediational effects using FPS and right amygdala activation similarity as mediators may be because these measures are typically measured less reliably (e.g., Shechner et al., 2015; see Table 2 in the Supplementary Materials). Mediation is less likely to be established when the mediator is measured less reliably (Lemmer & Gollwitzer, 2017). This issue is unfortunately a technical limitation related to collecting psychophysiological responses. Improved techniques for measuring psychophysiological responses may allow for establishing mediation with these responses and may increase the correspondence between these measures and subjective measures. However, the fact that we did already obtain some evidence for mediation of these measures by subjective fear ratings indicates that these two types of responses correspond more closely than sometimes assumed (e.g., LeDoux, 2014). In fact, we think that our data are most consistent with models of fear which propose that physiological and subjective fear responses constitute an integrated response to a threatening situation (e.g., Fanselow & Pennington, 2018). Future studies and technical improvements regarding psychophysiological measurement will clarify whether indeed more reliably measurement of psychophysiological responses results in a closer correspondence between these measures and subjective reports..

Our results also relate to the discussion about the validity of US expectancy ratings in fear conditioning (Boddez et al., 2012). That is, previously it has been argued that US expectancy ratings can be a valid and sensitive index of conditioned fear, although they may be sensitive to experimental demand effects. However, our results suggest that, to capture the variance of the effects of CS-US pairings, fear ratings seem to do better than US expectancy ratings. One conceptual reason why fear ratings outperform US expectancy ratings is that fear ratings capture both the expectation *and* evaluation of the US. This is particularly relevant when

participants do not find the US very aversive. In this case, participants may expect the US, but do not experience much fear because the US which they expect is not sufficiently aversive. In this situation, fear ratings would better capture their subjectively experienced fear. Another reason for the better performance of fear ratings may be due to the specific procedure used in these studies. That is, participants received very explicit instructions about the contingency between the CSs and the US. It is conceivable that participants take this information into account when providing their expectancy ratings, even though it may not reflect their actual expectancies. That is, participants may report similar expectancy for CS+P and CS+U, because the instructions clearly informed them that both CSs would be followed by an electric shock, but they may privately believe that the shock is more likely after CS+P than after CS+U. The effects for US expectancy may thus present an underestimation of the true difference in expectancy between CS+P and the CS+U. Fear ratings, on the other hand, may be less susceptible to these experimental demand effects of verbal instructions because the phrasing of the fear rating question ("How much fear did you experience while looking at this figure?") places more emphasis on participants' subjective experience. Thus, possibly fear ratings more closely reflect participants' actual expectancies than US expectancy ratings in procedures like the present one. Therefore, we argue that it may be relevant for future studies, especially studies investigating the effects of instructions on conditioned fear, to include fear ratings.

Finally, some limitations of our analyses should be acknowledged. One important limitation is the relatively small sample size of the different studies considered here, and thus the limited statistical power to detect mediational effects (see Montoya and Hayes, 2017). Particularly, mediation of the effect of CS-US on amygdala activation pattern similarity by fear ratings should be interpreted with caution given that it was only established with a 90% (and not

with a 95%) confidence interval. Ideally, the results of our analysis should be replicated in larger samples which include neural, psychophysiological (skin conductance and startle potentiation), and subjective measures (US expectancy and fear ratings). Nonetheless, the consistency of our results over studies indicates that the mediating role of fear ratings (and the lack of a mediating role of US expectancy ratings) is not a mere chance finding. Another limitation is the extent to which these mediation analyses allow us to draw conclusions about causality. That is, mediation analyses allow the evaluation of theoretical claims about causality, but cannot provide direct evidence for or against causality in itself (Kazdin, 2007). Indeed, significant results in a mediation analysis can be consistent with many different causal models (Fiedler, Schott, & Meiser, 2011). Nonetheless, our analyses do support the plausibility of certain models (e.g., mediation of the additive effects of CS-US pairings by fear ratings) and exclude other models (e.g., the additive effects of CS-US pairings can be explained by changes in US expectancy ratings). Finally, care should be taken generalizing our results to other procedures in which conditioned fear is investigated. That is, we considered the data of a very specific procedure in this report (i.e., the procedure developed by Raes et al., 2014). Hence, the results obtained here may not apply to other types of conditioning paradigms, such as uninstructed conditioning (for reviews see Lonsdorf et al., 2017; Mertens, Boddez, Sevenster, Engelhard, & De Houwer, 2018).

To conclude, a systematic re-analysis of the different studies from our labs that investigated the additive effects of CS-US pairings to contingency instructions demonstrates that differences in US expectancy ratings do not explain the additive effects of CS-US pairings on fear responses. In contrast, fear ratings do mediate the additive effects of CS-US pairings on US expectancy ratings and amygdala activation pattern similarity. These results demonstrate that changes in subjective fear are relevant to explain the additive effects of CS-US pairings. Our

results thus argue that fear ratings are a relevant measure to include in (instructed) fear conditioning studies and underline the possibility that US expectancy ratings may be affected by factors unrelated to fear, such as experimental demand effects, and may be unaffected by other factors important to fear, such as the US evaluation (e.g., Leer & Engelhard, 2015).

Acknowledgements

The research reported in this paper was funded by a NWO VICI grant (grant number: 453-15-005) awarded to Iris M. Engelhard, the Interuniversity Attraction Poles Program initiated by the Belgian Science Policy Office (IUAPVII/33) and by Ghent University Methusalem Grant BOF16/MET_V/002 awarded to Jan De Houwer.

References

Boddez, Y., Baeyens, F., Luyten, L., Vansteenwegen, D., Hermans, D., & Beckers, T. (2012). Rating data are underrated: Validity of US expectancy in human fear conditioning. *Journal of Behavior Therapy and Experimental Psychiatry*, *44*(2), 201–6. https://doi.org/10.1016/j.jbtep.2012.08.003

Braem, S., De Houwer, J., Demanet, J., Yuen, K. S. L., Kalisch, R., & Brass, M. (2017). Pattern analyses reveal separate experience-based fear memories in the human right amygdala. *The Journal of Neuroscience*, 908–17. https://doi.org/10.1523/JNEUROSCI.0908-17.2017

Davey, G. C. L. (1992). Classical conditioning and the acquisition of human fears and phobias: A review and synthesis of the literature. *Advances in Behaviour Research and Therapy*, *14*(1), 29–66. https://doi.org/10.1016/0146-6402(92)90010-L

Dawson, M. E., & Furedy, J. J. (1976). The Role of Awareness in Human Differential Autonomic Classical Conditioning: The Necessary-Gate Hypothesis. *Psychophysiology*, *13*(1), 50–53. https://doi.org/10.1111/j.1469-8986.1976.tb03336.x

Fanselow, M. S., & Pennington, Z. T. (2018). A return to the psychiatric dark ages with a two-system framework for fear. *Behaviour Research and Therapy*, *100*(October 2017), 24–29. https://doi.org/10.1016/j.brat.2017.10.012

Fiedler, K., Schott, M., & Meiser, T. (2011). What mediation analysis can (not) do. *Journal of Experimental Social Psychology*, *47*(6), 1231–1236. https://doi.org/10.1016/j.jesp.2011.05.007

Kazdin, A. E. (2007). Mediators and Mechanisms of Change in Psychotherapy Research. *Annual Review of Clinical Psychology*, *3*(1), 1–27. https://doi.org/10.1146/annurev.clinpsy.3.022806.091432

LeDoux, J. E. (2014). Coming to terms with fear. *Proceedings of the National Academy of Sciences of the United States of America, 111*(8), 2871–8. https://doi.org/10.1073/pnas.1400335111

Leer, A., & Engelhard, I. M. (2015). Countering Fear Renewal: Changes in the UCS Representation Generalize Across Contexts. *Behavior Therapy*, *46*(2), 272–282. https://doi.org/10.1016/j.beth.2014.09.012

Lemmer, G., & Gollwitzer, M. (2017). The "true" indirect effect won't (always) stand up: When and why reverse mediation testing fails. *Journal of Experimental Social Psychology*, *69*, 144–149. https://doi.org/10.1016/j.jesp.2016.05.002

Lonsdorf, T. B., Menz, M. M., Andreatta, M., Fullana, M. A., Golkar, A., Haaker, J., … Merz, C. J. (2017). Don't fear "fear conditioning": Methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. *Neuroscience & Biobehavioral Reviews*, *77*, 247–285. https://doi.org/10.1016/j.neubiorev.2017.02.026

Lovibond, P. F. (2011). Learning and anxiety: A cognitive perspective. In T. R. Schachtman & S. Reilly (Eds.), *Associative Learning and Conditioning: Human and non-human Applications* (pp. 104–120). New York: Oxford University Press.

Mertens, G., Boddez, Y., Sevenster, D., Engelhard, I. M., & De Houwer, J. (2018). A review on the effects of verbal instructions in human fear conditioning: Empirical findings, theoretical considerations, and future directions. *Biological Psychology*, *137*(October 2017), 49–64. https://doi.org/10.1016/j.biopsycho.2018.07.002

Mertens, G., Kuhn, M., Raes, A. K., Kalisch, R., De Houwer, J., & Lonsdorf, T. B. (2016). Fear expression and return of fear following threat instruction with or without direct contingency experience. *Cognition and Emotion*, *30*(5), 968–984.

https://doi.org/10.1080/02699931.2015.1038219

Mertens, G., Raes, A. K., & De Houwer, J. (2016). Can prepared fear conditioning result from verbal instructions? *Learning and Motivation*, *53*, 7–23. https://doi.org/10.1016/j.lmot.2015.11.001

Mineka, S., & Öhman, A. (2002a). Born to fear: Non-associative vs associative factors in the etiology of phobias. *Behaviour Research and Therapy*, *40*(2), 173–184. https://doi.org/10.1016/S0005-7967(01)00050-X

Mineka, S., & Öhman, A. (2002b). Phobias and preparedness: The selective, automatic, and encapsulated nature of fear. *Biological Psychiatry*, *52*(10), 927–937. https://doi.org/10.1016/S0006-3223(02)01669-4

Montoya, A. K., & Hayes, A. F. (2017). Two-condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods*, *22*(1), 6–27. https://doi.org/10.1037/met0000086

Öhman, A., & Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review*, *108*(3), 483–522. https://doi.org/10.1037//0033-295X.108.3.483

Olsson, A., & Phelps, E. a. (2007). Social learning of fear. *Nature Neuroscience*, *10*(9), 1095–102. https://doi.org/10.1038/nn1968

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 717–731. https://doi.org/10.3758/BF03206553

Rachman, S. (1977). The conditioning theory of fearacquisition: A critical examination. *Behaviour Research and Therapy*, *15*(5), 375–387. https://doi.org/10.1016/0005-

7967(77)90041-9

Raes, A. K., De Houwer, J., De Schryver, M., Brass, M., & Kalisch, R. (2014). Do CS-US pairings actually matter? A within-subject comparison of instructed fear conditioning with and without actual CS-US pairings. *PLoS ONE*, *9*(1). https://doi.org/10.1371/journal.pone.0084888

Reiss, S. (1980). Pavlovian conditioning and human fear: An expectancy model. *Behavior Therapy*, *11*(3), 380–396. https://doi.org/10.1016/S0005-7894(80)80054-2

Shechner, T., Britton, J. C., Ronkin, E. G., Jarcho, J. M., Mash, J. A., Michalska, K. J., … Pine, D. S. (2015). Fear conditioning and extinction in anxious and nonanxious youth and adults: Examining a novel developmentally appropriate fear-conditioning task. *Depression and Anxiety*, *32*(4), 277–288. https://doi.org/10.1002/da.22318

Supplementary material

Table 1. Pearson's correlation coefficients between the difference between CS+P and CS+U for the different measures and in the different datasets.

Measure	**Raes et al. (2014)**		**Mertens, Kuhn, et al. (2016)**			**Mertens, Raes, et al. (2016) Fear-irrelevant CSs**		**Mertens, Raes, et al. (2016) Fear-irrelevant CSs**		**Braem et al. (2017)**		
	1.	2.	1.	2.	3.	1.	2.	1.	2.	1.	2.	3.
Raes et al. (2014)												
1. US expectancy ratings	-	.638*										
2. Fear ratings		-										
Mertens, Kuhn, et al. (2016)												
1. US expectancy ratings			-	.751*	-.020							
2. Fear ratings				-	.197							
3. FPS					-							
Mertens, Raes, et al. (2016)												
Fear-irrelevant CSs												
1. US expectancy ratings						-	.700*					
2. Fear ratings							-					
Fear-relevant CSs												
1. US expectancy ratings								-	.612*			
2. Fear ratings									-			
Braem et al. (2017)												
1. US expectancy ratings										-	.929*	.394+
2. Fear ratings											-	.422+
3. Inter-CS activation similarity right amygdala												-

Note: * indicates $p < .05$; + indicates $p < .1$.

Table 2. Intraclass correlation coefficients and 95% confidence intervals for the different measures throughout the test phase.

Model	CS+U	CS+P
US expectancy ratings (N = 160)	.89** [.85–.92]	.86** [.82–.90]
Fear ratings (N = 160)	.89** [.86–.92]	.92** [89–94]
FPS (N = 35)	.58* [.25–.78]	.50* [.13–.73]
Inter-CS activation similarity right amygdala (N = 20)	.49* [-.07–.78]	.54* [.02–.80]

Note: ** indicates $p < .001$; * indicates $p < .05$.

Table 3. Model fit indices for the different mediation models.

Model	Chi-square	RMSEA	AIC	BIC
Model 1 (N = 160): M = Expec; O = Fear	$\chi2(1) < .001$ p = .998	< .01	1821.89	1846.49
Model 2 (N = 160): M = Fear; O = Expec	$\chi2(1) = .097$ p = .756	< .01	1826.08	1850.68
Model 3 (N = 35): M = Expec; O = Startle	$\chi2(1) = .082$ p = .960	< .01	527.47	539.91
Model 4 (N = 35): M = Startle; O = Expec	$\chi2(1) < .001$ p = .998	< .01	613.70	626.14
Model 5 (N = 35): M = Fear; O = Startle	$\chi2(1) = .035$ p = .852	< .01	509.80	522.24
Model 6 (N = 35): M = Startle; O = Fear	$\chi2(1) < .001$ p = .998	< .01	596.73	609.17
Model 7 (N = 20): M = Expec; O = Amyg	$\chi2(1) < .001$ p = .995	< .01	149.53	157.50
Model 8 (N = 20): M = Amyg; O = Expec	$\chi2(1) < .001$ p = 1	< .01	90.41	98.37
Model 9 (N = 20): M = Fear; O = Amyg	$\chi2(1) < .001$ p = .998	< .01	148.03	156.00
Model 10 (N = 20): M = Amyg; O = Expec	$\chi2(1) < .001$ p = 1	< .01	92.64	100.60

Note: For the analysis regarding the subjective measures we combined the data from all the different studies. Model fits were also excellent (RMSEA < .01) for the subjective ratings when evaluated for the individual studies. M = mediator; O = outcome; RMSEA = Root Mean Square Error of Approximation; AIC = Akaike information criterion; BIC = Bayesian information criterion.