The El Greco fallacy, this time with feeling: How (not) to measure group differences in emotional intensity

Vladimir Chituc (0000-0002-5316-6245) [1]
Brian J. Scholl (0000-0003-3610-0890) [1,2]

[1] Department of Psychology, Yale University, New Haven, CT, USA

[2] Wu-Tsai Institute, Yale University, New Haven, CT, USA

Running Head : El Greco with feeling

Address for correspondence : Vladimir Chituc
Department of Psychology
Yale University
Box 208047
New Haven, CT 06520-8047

Email : vladimir.chituc@yale.edu
Phone : 607-857-4166

Word Count : 3607
Version : 5/27/25 — In press, *Affective Science*

Abstract (212 words)

We all get angry. And some of us get angrier than others. But are such differences systematic across groups? Affective scientists often make claims about group differences in emotional intensity by comparing group averages on labeled (e.g. Likert) scales. For example, prominent research has used such scales to claim, perhaps counterintuitively, that women experience anger more intensely than men. Despite their ubiquity, these measures are susceptible to a subtle but notorious problem: the *El Greco fallacy*. (El Greco famously painted elongated figures. Could this be because he experienced a stretched-out world, due to a type of astigmatism? No, since then he would also have experienced a stretched-out canvas, and the effects would necessarily cancel out.) Here, across four experiments, we replicate the apparent finding with labeled scales that women experience anger more intensely than men. But we then go on to show that this finding reliably disappears when tested with a general Labeled Magnitude Scale (gLMS) — a psychophysical measure designed to detect such differences, avoiding the El Greco fallacy. This demonstrates how insights from sensory psychophysics can be usefully employed in affective science, and supports skepticism about purported group differences in emotional intensity based on labeled scales. The deep lesson for psychological measurement: the most obvious tools are not always the best.

Keywords

Emotion; Sex differences; Emotional intensity; gLMS; El Greco fallacy

Do you feel what I feel? Maybe not: as we are all aware, there are substantial differences in emotional experiences across people (and even across the same people at different times). Some differences may be qualitative, but they may more often involve emotional intensity (e.g. Diener, Larsen, et al., 1985). What elates me may only tickle you, and what enrages you may only bother me. This much seems obvious from our everyday experience of feeling and sharing emotions with one another, but affective science has sometimes gone further, suggesting that some such individual differences are *systematic*, varying across groups. For example, some researchers have claimed that women feel emotions more strongly than men (Davis et al., 2012; Diener, Sandvik, et al., 1985), or that people living in China experience emotions less intensely than people living in the West (Davis et al., 2012; Eid & Diener, 2001).

In the current short empirical report, we take no position on the truth or falsity of such claims (nor on any broader theoretical issues surrounding the nature of emotion, such as whether it is largely universal or socially constructed). The only claims about emotions that matter for our purposes are as follows: having an emotion *feels like something,* and this feeling can be stronger or weaker. Accordingly, our focus will be only on how such claims about group differences could and should (and couldn't and shouldn't) be measured. In particular, we will suggest that certain (almost universal) ways of measuring such claims may be flawed, for a famously subtle reason.

The *El Greco Fallacy*

Emotional experience is inherently private, making the *measurement* of emotion (or any sensation) a notoriously difficult (and historically controversial) enterprise. This isn't necessarily obvious, since it may seem at first blush like one could simply ask people to report details of their subjective experiences, including intensity. But report *how*? By far the most common answer involves *labeled scales* — e.g. reporting the intensity of felt anger on a Likert scale, anchored with labels from "not at all angry" to "very angry". Though pleasingly direct, this

method masks a subtle and fascinating problem: differences in underlying experiences may also create differences in the use of the scale.

The most infamous example of this may be a proposed explanation for the distinctive style of the Spanish Renaissance painter El Greco. El Greco famously painted figures that were exceptionally (and even oddly) *elongated*. Why? One famous proposal implicated El Greco's perception of the world: if he perceived the world as vertically stretched due to a type of astigmatism, then perhaps he just painted what he saw. On reflection, however, this explanation could not possibly work, since such an astigmatism would have caused El Greco to also experience a stretched-out canvas, such that the effects would cancel out (for a review, see Firestone, 2013). Thinking otherwise has come to be known as the *El Greco fallacy*, and this error in reasoning seems as stubborn as it is seductive. For example, modern research has argued that several alleged 'top-down' effects of cognition on perception (e.g. the suggestion that reflecting on immoral actions makes the world appear darker; Banerjee et al., 2012) could not possibly be true, because they inadvertently commit the fallacy (see Firestone & Scholl, 2014, 2016). (And the fallacy may even persist in modern art appreciation, as when a team of otolaryngologists recently proposed that vestibular migraines could explain why Van Gogh painted with a tilt to the left; Dasgupta et al., 2022; cf. Huntley, 2022).

The Current Studies

In the present project, we consider the El Greco fallacy in the context of alleged group differences in emotional experience. The theoretical motivation for claims of such differences can vary. For example, perhaps women experience emotions like fear more intensely as an adaptive strategy in response to greater physical vulnerability (e.g. Benenson et al., 2022); or perhaps cross-cultural differences in emotion stem from differences in cultural norms surrounding the *expression* of emotion, since that would encourage downregulation or reappraisal strategies to suppress the experience of strong emotions (Eid & Diener, 2001). While such claims could be true, it seems challenging to substantiate them with labeled scales —

because such scales (like El Greco's canvas) could be distorted in the same way as the emotional experience purportedly measured by those scales. Consider the claim that women experience *anger* more intensely than men (Simon & Nath, 2004). This could not be readily captured on a labeled scale, since the anchors themselves — e.g. "not at all intense" or "very intense" — would naturally be interpreted differently, due to the very same putative differences in emotional intensity (Fig. 1).

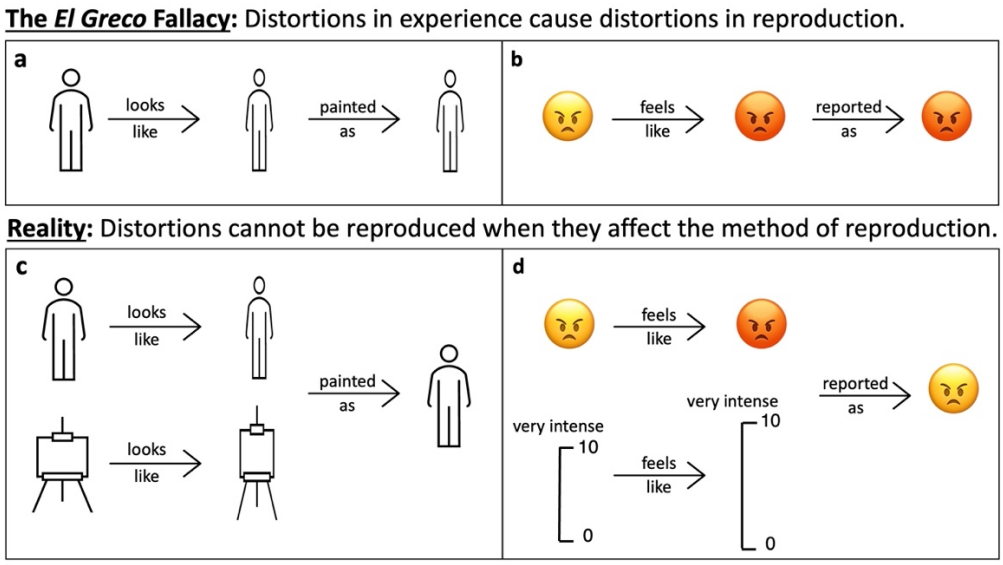

Fig. 1.
Illustration of the *El Greco fallacy*. The distortion of a stimulus is assumed to be reflected in the reproduction of that stimulus, as depicted in the context of vertical elongation captured by a painting (a) and more intensely felt anger reported on a scale (b). In reality, a perceptual distortion should be undetectable if it also affects the method of reproducing that distortion. Vertical elongation of a figure cannot be captured by a painting of that figure, since the canvas would also be perceived as vertically elongated in kind (c). Just so, a magnified experience of anger cannot be captured on a labeled scale of intensity, since the meaning of "very intense" anger itself would be magnified in kind (d).

This point may be subtle, but it has been well understood in the study of sensory psychophysics. Consider, for example, *supertasters* — who experience tastes as much more intense than do most people (something true, incidentally, of both authors). Such differences are very real, but can often go undetected by labeled scales (for a review, see Bartoshuk et al., 2005).

Normal tasters and supertasters may both describe a soda as sweet, and may even provide numerically identical ratings on a scale ranging from "not at all sweet" to "very sweet". But the actual anchoring experience of "very sweet" itself differs dramatically across such groups, such that (despite identical ratings!) supertasters may experience soda to be dramatically more sweet than would normal tasters (Bartoshuk, 2014; Bartoshuk et al., 2003).

Here, we directly apply this logic to the claim that women experience anger more intensely than men (Simon & Nath, 2004). We chose this target finding because it is widely cited (over 950 times as of this writing), it is replicable (e.g. Brebner, 2003; Fischer & Roseman, 2007), and it runs counter to the stereotype that anger is more typically associated with men's emotions (e.g. Kelly & Hutson-Comeaux, 1999).[1]

We first aim, in Experiment 1a, to conceptually replicate the apparent finding that women experience more intense anger, using the same measure employed in previous work: a 10-point Likert scale, ranging from "not at all intense" to "very intense" — which is susceptible in principle to the El Greco fallacy. Then, in Experiment 1b, we aim to make the same comparison using a scale that was explicitly developed to capture true underlying differences in subjective sensory experience: the gLMS (general Labeled Magnitude Scale; Bartoshuk, Duffy, Green, et al., 2004; Hayes et al., 2013). Importantly, the gLMS employs a benchmark systematically unrelated to the intensity of any one sensation, making it immune to the El Greco fallacy (at least in this context). More specifically, the gLMS asks subjects to rate their experience relative to "the strongest imaginable sensation" of any type, rather than to an extreme intensity of the target experience itself. In Experiments 2a and 2b, we then provide a further conceptual replication of both the Likert and gLMS experiments, using a subtly different variety of gLMS measurement.

[1] The relevant question here is thus whether the same anger-inducing stimulus corresponds to different underlying *experiences* of anger. Note that this is independent of whether or not women encounter more anger-inducing stimuli (in either frequency or intensity) than do men.

Experiments 1a-1b

Based on the logic of the El Greco fallacy, we predicted that the apparent difference in emotional intensity — with women experiencing anger more intensely compared to men — would replicate with the typical Likert scale measurement (in Experiment 1a), but may disappear with the more careful gLMS measurement (in Experiment 1b).

Method

All hypotheses, sample sizes, methods, and analyses were preregistered before data collection began (see https://aspredicted.org/MM4_KNV) — with the only deviation from preregistration being the exclusion of subjects who inadvertently participated in multiple experiments from this same project (though retaining these subjects produces qualitatively identical results).

Subjects

Via the Prolific survey platform (Palan & Schitter, 2018), we recruited a convenience sample of 1,000 subjects for each experiment (500 men and 500 women, based on demographic information that they had previously reported to Prolific; n=4,000). Though we preregistered no exclusions, there was a brief period when more than one experiment reported here was recruiting concurrently, inadvertently leading 525 subjects to complete more than one of them. In such cases, we retained only the first experiment completed by that subject, and we recruited additional subjects until we reached our preregistered sample size, which had 99% power to detect an effect as large as the one from the initial report of such effects (d=0.3, as calculated from the sample size, means, and SDs from Table B1 of Simon and Nath, 2004, p. 1172).

Stimuli and Procedure

All testing was conducted using the Qualtrics survey platform. Subjects were first instructed to think about the anger they felt the last time someone insulted them — describing what happened in a free-response box, per the instructions: “In only a few words, describe what

was happening that made you feel this way."[2] (We intentionally asked for an 'everyday' instance of anger in this context — rather than an emotional extreme, such as "the angriest you've ever felt" — [1] to avoid possible ceiling effects in the Likert data, [2] to make it more relevant to everyday life and most emotion research, and [3] because the initial research of Simon and Nath did not focus on extremes.) They were then subsequently asked to report the intensity of the anger they felt using one of two measurement methods — a Likert scale (Experiment 1a) or the gLMS (Experiment 1b). For the Likert scale, following the initial report of such effects (Simon & Nath, 2004), subjects simply clicked on one of 10 visible numbers, ranging from 1 (anchored with the label "not at all intense") to 10 (anchored with the label "very intense"). For the gLMS, subjects clicked and dragged a slider along a vertical line, ranging from 0 (anchored with the label "no sensation") to 100 (anchored with the label "strongest imaginable sensation"), with other labels (e.g. "moderate" and "very strong") quasi-logarithmically spaced across the length of the scale (Fig. 2).

[2] The original report of such differences (Simon & Nath, 2004) used data taken from the emotion module of the 1996 General Social Survey (GSS), which asked respondents multi-stage questions involving the frequency, intensity, and duration for 19 different emotions. Since we were only interested the intensity of anger, we elicited ratings in a more simplified way.

Fig. 2.
The general Labeled Magnitude Scale (gLMS) as displayed to subjects. Since group differences in the intensity of anger are unrelated to the strongest sensation one can imagine, the gLMS is not susceptible to the El Greco fallacy. It is for this reason that the gLMS has been widely used to detect (or rule out) group differences or changes in sensory experience, including for taste (Bartoshuk, Duffy, Green, et al., 2004), pain (Bartoshuk, Duffy, Chapo, et al., 2004; Čeko et al., 2022), and smell (Petrova et al., 2008).

The instructions explaining this scale (adapted from Hayes et al., 2013) were as follows:

> *The scale you will use today captures the entire range of how intense experiences can possibly be. The top of the scale (100) is the strongest sensation you could even imagine, which should be the maximum amount of intensity possible for an experience. The bottom of the scale (0) is the absence of sensation, which should be the minimum amount of intensity possible for an experience.*
>
> *Between those two extremes, you should be able to rate the intensity of every other experience you have ever had or could even imagine having. If there's a sensation that would fall outside either end of the scale, then that sensation is what should be the endpoint, instead.*
>
> *This is very important: the endpoints of the scale represent the minimum and maximum level of intensity possible for an experience, so you should be able to assign any other experience a number between 0 and 100.*

Results

Per our preregistered analyses, we first tested for a difference between men and women on the reported intensity of anger using a Likert scale. Replicating the original result (Simon & Nath, 2004), women reported more intense anger than did men (7.0 [SD 2.0] vs. 6.7 [SD 2.1]; Welch two samples $t(996.4)=2.60$, $p=.009$, $d=0.16$ 95% CI = [0.04, 0.29]; this and all subsequent tests are 2-tailed). We next tested whether this difference would also appear using the gLMS. As predicted, it did not: women and men reported similarly intense anger (44.2 [*SD* 25.0] vs. 42.5 [*SD* 25.1]; $t(998.0)=1.04$, $p=.301$, $d=0.07$ [-0.06, 0.19]). We next tested whether the pattern of results produced by Likert ratings differed significantly from that produced by gLMS ratings. After normalizing each measure (z-scoring the raw Likert ratings and the logarithm of the gLMS ratings [incremented by one to avoid undefined values]), a 2x2 Between-subjects ANOVA (measure: Likert vs. gLMS; sex: male vs. female) revealed a significant interaction ($F(3, 1996)=3.29$, $p=.020$; $\eta^2=.005$ [0.001, 0.012]).

Experiments 2a-2b

As typically used, the gLMS includes a standard battery of 15 practice stimuli to orient subjects to the use of the scale (e.g. Bartoshuk et al., 2003; Hayes et al., 2013). Because Experiment 1b did not employ these practice trials, we replicated the results of both experiments with the addition of this battery in Experiment 2b.

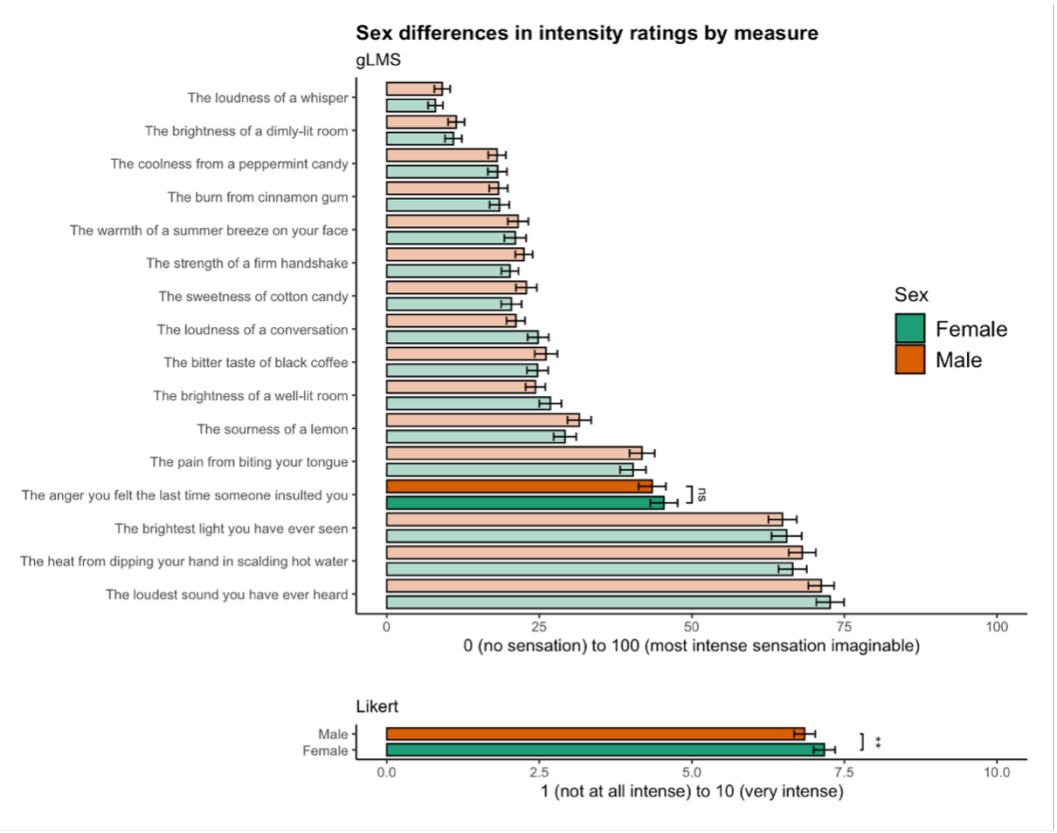

Fig. 3.
Mean ratings from Experiment 2a (bottom) and 2b (top), split by sex. Anger intensity is presented in the solid bars, and the battery of practice questions are presented in faded bars. Error bars represent 95% confidence intervals. ** indicates differences of p=.01.

Method

Experiment 2a was a direct replication of Experiment 1a. Experiment 2b was identical to Experiment 1b except that before rating the intensity of their anger, subjects first completed 15 practice trials, using the gLMS with sample items relating to other forms of experience (including “The strength of a firm handshake”, “The warmth of a summer breeze on your face”, and “The brightest light you have ever seen”; for full details, see Bartoshuk et al., 2003; Hayes et al., 2013). For each experiment, we recruited a new sample of 1,000 subjects (500 men, and 500 women), post-exclusions (as described above), with this preregistered sample size chosen to match that of Experiments 1a and 1b.

Results

Replicating the original result (Simon & Nath, 2004), women again reported more intense anger than did men, when assessed with a Likert scale (7.2 [SD 2.0] vs. 6.8 [SD 2.0]; $t(997.3)=2.57$, $p=.010$, $d=0.16$ [0.04, 0.29]). But when tested with the gLMS, this difference again disappeared, with women and men reporting similarly intense anger (45.4 [*SD* 25.5] vs. 43.5 [*SD* 25.5]; $t(998)=1.20$, $p=.23$, $d=0.08$ [-0.06, 0.19]). After normalization, a 2x2 Between-subjects ANOVA again found a statistically significant interaction ($F(3, 1996)=2.66$, $p=.047$; $\eta^2=.004$ [0.001, 0.010]).

Next, we preregistered an exploratory analysis to determine whether there were any overall sex differences in the use of the gLMS and, if so, to correct for them (Fig. 3). To determine whether there were such sex differences, we conducted a 2 (Between-subjects factor: sex) x 15 (Within-subjects factor: practice question) mixed-measures ANOVA. This revealed a significant interaction between sex and practice question ($F(7.9, 7792)=2.94$, $p=.003$; $\eta^2=.003$ [0.001, 0.005]). To correct for these differences, we then normalized all 16 gLMS responses (15 practice questions and the intensity of anger) separately for both men and women. First, we tested whether these corrected gLMS ratings would reveal a sex difference in anger, such that women might report higher ratings than did men, and they did not: the results were qualitatively identical ($t(989.3)=1.55$, $p=.122$, $d=0.10$ [-0.03, 0.22]). Next, we centered the gLMS means around 0, recombined it with the Likert data, and tested again for the interaction, which remained significant ($F(3, 1996)=3.47$, $p=.016$; $\eta^2=.005$ [0.001, 0.012]).

Finally, we report one additional exploratory analysis to support the overall robustness of our results. To achieve greater statistical power, we combined the datasets for Experiments 1b and 2b. Though this sample is sufficiently powered to detect an effect size of only $d=0.14$ with 99% power (which we note is even smaller than the one obtained by Likert ratings in either Experiment 1a [$d=0.17$] or Experiment 1b [$d=0.16$]) and less than half the size of the original

finding [d=0.3], we again found no difference in anger between gLMS ratings provided by men and women; $t(1998)=1.58$, $p=.114$, $d=0.07$ [-0.02, 0.16].

General Discussion

Claims about group differences in emotional intensity are widespread in affective science, similar to the particular claim tested here — that women experience anger more intensely compared to men. And indeed, we ourselves replicated this very result twice in the current project (in Experiments 1a and 2a). We nevertheless remain skeptical about this and similar claims, both in principle and in practice, due to the potential (and unacknowledged) influence of the El Greco fallacy: El Greco's elongated paintings could not be explained by an elongated perception of the world, since he would then also see the canvas as elongated, and these distortions would cancel out in the paintings themselves. Analogously, we should be unable to detect that women experience anger more intensely than do men using labeled (e.g. Likert) scales, since such differences in intensity would affect the interpretations of the scale's extremities as well, such that these differences would cancel out.

As predicted by this interpretation, the putative sex difference in anger disappeared when tested with the more careful gLMS measure that was developed precisely to detect actual differences in experience. This pattern seemed especially robust. *First*, it was replicated multiple times (across Experiments 1 and 2). *Second,* the difference itself was also apparent in statistically reliable interactions that directly compared the two types of measures. And *third*, the gLMS failed to find an effect even when testing with extremely high power (combining data across Experiments 1b and 2b).

But Why?

The current results suggest that women may not in fact experience anger more intensely than men. But if so, why do such putative differences reliably appear when measured with

labeled scales? An obvious limitation of the present study is that it cannot answer this question. According the logic of the El Greco fallacy, however, this question does not have to be answered in order for us to know that the effect cannot reflect *actual* differences in emotional experience, per se. (Similarly, we can rule out astigmatism before discovering the true reason that El Greco painted elongated figures; see Firestone & Scholl, 2014.)

Nevertheless, we might speculate about this by noting that these results are consistent with a number of possible explanations involving gender norms or stereotypes (e.g. Huntsinger & Raoul, 2022). One possibility, in particular, is that this difference may (somewhat ironically) be caused by the widespread folk assumption that women experience anger *less* intensely than do men. Consider the following seemingly-paradoxical pattern of results from a study on stereotyping: when judging the financial success of different men and women, subjects reported that the men earned higher salaries than the women, yet they also reported that the women were more financially successful than the men (as rated on a 7-point Likert scale; Biernat et al., 1991). The authors interpret these results as reflecting gendered stereotypes and expectations involving women's financial success: if women are held to a lower standard of financial success, then a phrase like "very financially successful" would correspond to a lower salary.

In an analogous way, an apparent gender difference in the experience of anger (as seemingly revealed by labeled scales) may instead be reflecting gendered stereotypes involving the levels of anger that are either expected or appropriate: men may be interpreting the phrase "very angry" to mean "very angry *for a (typical) man*," while women may be interpreting it to mean "very angry *for a (typical) woman*". Thus, women's anger may be rated more highly (despite identical experience) in the same way that they may also be rated as more financially successful (despite lower salaries). Of course, this is only one of many such plausible accounts, and the explanation for this illusory difference in Likert ratings remains an open question that future work may wish to explore. And of course, it may still be true after all that women experience anger more intensely than do men. As noted above, we take no position on that

claim here; our only (important!) point is that the Likert-based experiments do not (and perhaps cannot) establish that claim.

Theoretical and Methodological Lessons, Beyond Anger

Though the empirical scope of this short report was intentionally narrow, the theoretical and methodological lessons are much broader. Theoretically, this work suggests that we should be skeptical of any purported group difference in emotional experience that is established using a labeled scale. And indeed, such skepticism may help make sense of confusing (if not contradictory) patterns of results, sometimes even obtained within a single study. For example, it has long been recognized that emotion ratings on labeled scales correlate only weakly (if at all) with physiological measures of emotion (Mauss et al., 2005). This has led some researchers to suggest abandoning certain physiological measures (Poláčková Šolcová & Lačev, 2017), but we suggest that the opposite conclusion may also be reasonable: subjective self-reports of emotion have been so poorly aligned with physiological measures because subjective self-reports of emotion have (so-far) been poorly aligned with emotional experience itself!

Methodologically, the central lesson is that this problem is extremely straightforward to solve. Taking a cue from work on group differences in sensory psychophysics, work exploring group differences in emotional intensity should be mindful of the El Greco fallacy by using methods to which this fallacy does not apply — such as the gLMS.[3] Such changes would be relatively simple to implement, as exemplified by the current project. They might nevertheless be highly consequential, if only because this insight from sensory psychophysics has almost never previously been adopted in this research community — which has instead relied almost universally on labeled scales. Indeed, we are aware of only a single instance of a different psychophysical method (magnitude estimation) being used in the context of emotion research

[3] As the gLMS may be somewhat awkward to implement in common platforms like Qualtrics (e.g. requiring custom code to space labels at unequal intervals), there are more streamlined alternatives such as the gVAS (general Visual Analog Scale; Hayes et al., 2013) which can also be used for this purpose.

(Hsee & Tang, 2007), and we are unaware of *any* instances of psychophysical scales being used to measure emotional intensity. We believe that a broader application of these measures may greatly aid affective scientists in more accurately measuring emotional experience — just as has been the case with the measurement of other types of sensations.

Statements and Declarations

Ethics approval: This research complies with the Declaration of Helsinki (2023), and received approval from the Yale University Institutional Review Board (ID: 2000022385).

Consent to participate: All subjects provided informed consent prior to taking part in this research.

Consent to publish: N/A

Availability of data and materials: All primary data for each experiment can be viewed in the Supplemental Material available online. The hypotheses, methods, analysis plans, and sample sizes are preregistered at https://aspredicted.org/MM4_KNV.

Code availability: Analysis scripts can be found in the Supplemental Material available online.

Competing interests: None.

Funding: This work was supported by a National Science Foundation Graduate Research Fellowship awarded to VC [grant number: DGE1752134].

Author's contributions: VC conceptualized the study, collected the data, and performed the statistical analyses and data visualizations under the supervision of BJS. Both authors designed the methodology and jointly wrote the original draft, as well as later revising and editing.

Acknowledgements: For helpful conversation and/or comments on previous drafts, we thank Molly Crockett, Maria Gendron, Wendy Berry Mendes, and the members of the Yale Affective Science and Culture Lab and the Yale Perception & Cognition Lab.

References

Banerjee, P., Chatterjee, P., & Sinha, J. (2012). Is It light or dark? Recalling moral behavior changes perception of brightness. *Psychological Science*, *23*(4), 407–409. https://doi.org/10.1177/0956797611432497

Bartoshuk, L. M. (2014). The measurement of pleasure and pain. *Perspectives on Psychological Science*, *9*(1), 91–93. https://doi.org/10.1177/1745691613512660

Bartoshuk, L. M., Duffy, V. B., Chapo, A. K., Fast, K., Yiee, J. H., Hoffman, H. J., Ko, C.-W., & Snyder, D. J. (2004). From psychophysics to the clinic: Missteps and advances. *Food Quality and Preference*, *15*(7–8), 617–632. https://doi.org/10.1016/j.foodqual.2004.05.007

Bartoshuk, L. M., Duffy, V. B., Fast, K., Green, B. G., Prutkin, J., & Snyder, D. J. (2003). Labeled scales (e.g., category, Likert, VAS) and invalid across-group comparisons: What we have learned from genetic variation in taste. *Food Quality and Preference*, *14*(2), 125–138.

Bartoshuk, L. M., Duffy, V. B., Green, B. G., Hoffman, H. J., Ko, C.-W., Lucchina, L. A., Marks, L. E., Snyder, D. J., & Weiffenbach, J. M. (2004). Valid across-group comparisons with labeled scales: The gLMS versus magnitude matching. *Physiology & Behavior*, *82*(1), 109–114. https://doi.org/10.1016/j.physbeh.2004.02.033

Bartoshuk, L. M., Fast, K., & Snyder, D. J. (2005). Differences in our sensory worlds: Invalid comparisons with labeled scales. *Current Directions in Psychological Science*, *14*(3), 122–125. https://doi.org/10.1111/j.0963-7214.2005.00346.x

Benenson, J. F., Webb, C. E., & Wrangham, R. W. (2022). Self-protection as an adaptive female strategy. *Behavioral and Brain Sciences*, *45*, e128. https://doi.org/10.1017/S0140525X21002417

Biernat, M., Manis, M., & Nelson, T. E. (1991). Stereotypes and standards of judgment. *Journal of Personality and Social Psychology*, *60*(4), 485–499. https://doi.org/10.1037/0022-3514.60.4.485

Brebner, J. (2003). Gender and emotions. *Personality and Individual Differences*, *34*(3), 387–394. https://doi.org/10.1016/S0191-8869(02)00059-4

Čeko, M., Kragel, P. A., Woo, C.-W., López-Solà, M., & Wager, T. D. (2022). Common and stimulus-type-specific brain representations of negative affect. *Nature Neuroscience*, *25*(6), 760–770. https://doi.org/10.1038/s41593-022-01082-w

Dasgupta, S., Vanspauwen, R., Guneri, E. A., & Mandala, M. (2022). Vincent Van Gogh and the elusive diagnosis of vestibular migraine. *Medical Hypotheses*, *159*, 110747. https://doi.org/10.1016/j.mehy.2021.110747

Davis, E., Greenberger, E., Charles, S., Chen, C., Zhao, L., & Dong, Q. (2012). Emotion experience and regulation in China and the United States: How do culture and gender shape emotion responding? *International Journal of Psychology*, *47*(3), 230–239. https://doi.org/10.1080/00207594.2011.626043

Diener, E., Larsen, R. J., Levine, S., & Emmons, R. A. (1985). Intensity and frequency: Dimensions underlying positive and negative affect. *Journal of Personality and Social Psychology*, *48*(5), 1253–1265.

Diener, E., Sandvik, E., & Larsen, R. J. (1985). Age and sex effects for emotional intensity. *Developmental Psychology*, *21*(3), 542–546. https://doi.org/10.1037/0012-1649.21.3.542

Eid, M., & Diener, E. (2001). Norms for experiencing emotions in different cultures: Inter- and intranational differences. *Journal of Personality and Social Psychology*, *81*(5), 869–885. https://doi.org/10.1037/0022-3514.81.5.869

Firestone, C. (2013). On the origin and status of the "El Greco fallacy." *Perception*, *42*(6), 672–674. https://doi.org/10.1068/p7488

Firestone, C., & Scholl, B. J. (2014). “Top-down” effects where none should be found: The El Greco fallacy in perception research. *Psychological Science*, *25*(1), 38–46. https://doi.org/10.1177/0956797613485092

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for “top-down” effects. *Behavioral and Brain Sciences*, *39*, e229. https://doi.org/10.1017/S0140525X15000965

Fischer, A. H., & Roseman, I. J. (2007). Beat them or ban them: The characteristics and social functions of anger and contempt. *Journal of Personality and Social Psychology*, *93*(1), 103–115. https://doi.org/10.1037/0022-3514.93.1.103

Hayes, J. E., Allen, A. L., & Bennett, S. M. (2013). Direct comparison of the generalized visual analog scale (gVAS) and general labeled magnitude scale (gLMS). *Food Quality and Preference*, *28*(1), 36–44. https://doi.org/10.1016/j.foodqual.2012.07.012

Hsee, C. K., & Tang, J. N. (2007). Sun and water: On a modulus-based measurement of happiness. *Emotion*, *7*(1), 213–218. https://doi.org/10.1037/1528-3542.7.1.213

Huntley, J. S. (2022). Van Gogh, lateral tilt, and the El Greco fallacy. *Medical Hypotheses*, *163*, 110861. https://doi.org/10.1016/j.mehy.2022.110861

Huntsinger, J. R., & Raoul, A. (2022). Only as a last resort: Sociocultural differences between women and men explain women’s heightened reaction to threat, not evolutionary principles. *Behavioral and Brain Sciences*, *45*, e140. https://doi.org/10.1017/S0140525X22000516

Kelly, J. R., & Hutson-Comeaux, S. L. (1999). Gender-emotion stereotypes are context specific. *Sex Roles*, *40*(1/2), 107–120. https://doi.org/10.1023/A:1018834501996

Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. *Emotion*, *5*(2), 175–190. https://doi.org/10.1037/1528-3542.5.2.175

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004

Petrova, M., Diamond, J., Schuster, B., & Dalton, P. (2008). Evaluation of trigeminal sensitivity to ammonia in asthmatics and healthy human volunteers. *Inhalation Toxicology*, *20*(12), 1085–1092. https://doi.org/10.1080/08958370802120396

Simon, R. W., & Nath, L. E. (2004). Gender and emotion in the United States: Do men and women differ in self-reports of feelings and expressive behavior? *American Journal of Sociology*, *109*(5), 1137–1176. https://doi.org/10.1086/382111