Journal of Pediatric Psychology, 2019, 1–12
doi: 10.1093/jpepsy/jsy104
Original Research Article

Gender Bias in Pediatric Pain Assessment

Brian D. Earp,[1] MSc, M.Phil, Joshua T. Monrad,[1] Marianne LaFrance,[1] PhD, John A. Bargh,[1] PhD, Lindsey L. Cohen,[2] PhD, and Jennifer A. Richeson[1] PhD

[1]Yale University and [2]Georgia State University

All correspondence concerning this article should be addressed to Brian D. Earp, Associate Director, Yale-Hastings Program in Ethics and Health Policy, Yale University and The Hastings Center, 2 Hillhouse Avenue, New Haven, CT 06511, USA. E-mail: brian.earp@yale.edu

Received June 15, 2018; revisions received November 19, 2018; accepted November 26, 2018

Abstract

Objective Accurate assessment of pain is central to diagnosis and treatment in healthcare, especially in pediatrics. However, few studies have examined potential biases in adult observer ratings of children's pain. Cohen, Cobb, & Martin (2014. Gender biases in adult ratings of pediatric pain. *Children's Health Care*, *43*, 87–95) reported that adult participants rated a child undergoing a medical procedure as feeling more pain when the child was described as a boy as compared to a girl, suggesting a possible gender bias. To confirm, clarify, and extend this finding, we conducted a replication experiment and follow-up study examining the role of explicit gender stereotypes in shaping such asymmetric judgments. **Methods** In an independent, pre-registered, direct replication and extension study with open data and materials (https://osf.io/t73c4/), we showed participants the same video from Cohen et al. (2014), with the child described as a boy or a girl depending on condition. We then asked adults to rate how much pain the child experienced and displayed, how typical the child was in these respects, and how much they agreed with explicit gender stereotypes concerning pain response in boys versus girls. **Results** Similar to Cohen et al. (2014), but with a larger and more demographically diverse sample, we found that the "boy" was rated as experiencing more pain than the "girl" despite identical clinical circumstances and identical pain behavior a cross conditions. Controlling for explicit gender stereotypes eliminated the effect. **Conclusions** Explicit gender stereotypes—for example, that boys are more stoic or girls are more emotive—may bias adult assessment of children's pain.

Key words: gender; pain; pediatric; stereotypes.

Introduction

Across healthcare settings and in everyday life, accurate assessments of another's pain are critical for guiding appropriate responses. Yet pain is a private experience and as such is not directly accessible to outside observers. Instead, an inference must be drawn from behavioral and situational cues, which are often ambiguous and may therefore be interpreted in a biased or otherwise distorted manner (Higgins, 1996). In the case of children, the ability to verbalize pain in a reliable way may not yet be fully developed, creating more room for ambiguity. Children also lack full autonomy, and must therefore often rely on others to meet their pain control needs (see Earp, in press). In pediatrics, for example, parents and medical staff are often responsible for evaluating and responding to children's pain, especially when the child is very young (Cohen et al., 2008). In such cases, adult inferences about the child's subjective experience of pain are central to understanding pathology and determining a suitable intervention. It is therefore crucial to identify any biasing factors that may influence such third-party pain assessments.

In this article, we focus on potential biases related to the perceived gender of a child. We define perceived

gender as the gender an observer assumes of a target individual, whether correctly or incorrectly, whereas actual gender refers to the gender by which the individual self-identifies (e.g., male, female, genderqueer, non-binary, transgender; for discussion, see Earp, 2016a; Harrison, Grant, & Herman, 2012; Richards et al., 2016). Neither concept should be conflated with a target individual's sex, a term used in pain research to refer to "physiological differences between males and females, [including] genetics, anatomy, and hormonal and immune functioning" (Keogh, 2018, p. 434). In other words, sex refers to certain bodily attributes, whereas gender is used in pain research to refer to the wider cultural or psychosocial significance of sex. This may involve constructs such as masculinity and femininity, understood as a cluster of behavioral norms, characteristics, and social expectations that are stereotypically associated with persons presumed to be of male or female sex, respectively (Keogh, 2018).

For the sake of simplicity and continuity with previous research, we will be using the binary terms "boy" and "girl" to refer to the target child's gender in this study. This is also consistent with the categorical way in which gender is automatically processed in, for example, face perception (Freeman, Rule, Adams, & Ambady, 2010). However, the question of how the pain experiences of genderqueer or non-binary children, including (some) trans or intersex children, are interpreted by adults is of great practical, theoretical, and ethical significance, and we hope to investigate such cases in future research (see Steinfeld & Earp, 2017).

Some studies looking at pediatric pain and its assessment have taken child sex and/or gender into account. According to a recent meta-analysis, male and female children usually give similar self-report ratings for pain intensity, pain threshold, pain tolerance, and pain affect in the cold pressor task, with no statistically or clinically significant differences prior to puberty (Boerner, Birnie, Caes, Schinkel, & Chambers, 2014). Other pain stimuli, such as experimental heat pain and pressure pain, have been less well-studied in children, and the existing evidence is insufficiently robust to draw strong conclusions about the presence or absence of a sex-based difference on such measures (Boerner et al., 2014).

With respect to pain treatment, there is little reliable information concerning possible disparities in children as a function of either sex or gender prior to adolescence; this has been identified as a priority for future research (Musey et al., 2014). In contrast, the literature on adults has identified numerous gender differences—potentially corresponding to sex, although this distinction is typically neglected in the relevant studies—with women receiving less adequate pain medication compared with men, having a lower probability of admission to intensive care units, and a higher likelihood of being denied additional diagnostic procedures in response to complaints of pain (see Bernardes, Keogh, & Lima, 2008 for a review including exceptions and contrary findings). Where such differences occur, they may plausibly be explained by a number of factors including some that correspond at least in part to genuine differences between adult males and females: for example, differences in average or typical pain responses along various biopsychosocial dimensions (Racine et al., 2012; Sorge & Strath, 2018; see also work on the social communication model of pain, e.g., Craig, 2018).

However, prior expectations or biases concerning culturally defined male and female gender roles—which do not necessarily apply at the individual level—may also be a part of the explanation (Myers, Riley, & Robinson, 2003; Robinson & Wise, 2003; Robinson et al., 2001; Wise, Price, Myers, Heft, & Robinson, 2002). Hoffmann and Tarzian (2001) argue that "the subjective nature of pain requires health-care providers to view the patient as a credible reporter, and stereotypes or assumptions about behavior in such circumstances (oversensitivity, complaining, stoicism) add to the likelihood of undertreatment of some groups and overtreatment of others" (p. 20). Specifically, insofar as women are assumed to be oversensitive and more emotionally expressive (Barrett & Bliss-Moreau, 2009; Hutson-Comeaux & Kelly, 2002), their requests for pain treatment may be taken less seriously, whereas if men are generally seen as more stoic and hence reluctant to report pain, their requests may be taken more seriously (Bernardes & Lima, 2011; Schäfer, Prkachin, Kaseweter, & Williams, 2016).

Do similar stereotypes influence the assessment of children's pain? To date, this possibility has received little empirical attention. Among the studies that do exist, the primary focus has often been on the gender of the adult observer (e.g., comparing pain ratings of mothers and fathers; Rosenbloom et al., 2011; Vervoort, Huguet, Verhoeven, & Goubert, 2011). In contrast, when the child's gender has been the focus, the child's actual and perceived gender are typically confounded, because the child's gender is known to the adult (e.g., Moon et al., 2008). Thus, any differences in pain ratings based on sex or gender could reflect either true differences in the pain response of boys when compared with girls (Boerner et al., 2014), or biased interpretations of such responses based on stereotypes held by the observer (or some combination of both). To tease these possibilities apart, an experimental design is required that varies the perceived gender of the child, while holding all else constant (Hirsh, Alqudah, Stutts, & Robinson, 2008).

To our knowledge, only one study has manipulated a child's perceived gender while controlling the source and display of pain. Cohen, Cobb, and Martin (2014) asked 183 undergraduate students, of whom 154 were female (i.e., 85% of the sample), to rate the pain of a 5-year-old child dressed in gender-neutral clothing who was undergoing a medical procedure as seen in a video. The same video was presented to participants regardless of condition, with the child described as a boy, "Samuel," in one condition, and as a girl, "Samantha," in the other condition. Surprisingly, the authors found that the "boy" was rated as experiencing *more* pain than the "girl" despite identical clinical circumstances and an identical behavioral display of pain.

This result is surprising because it appears to conflict with the common cultural belief that girls are more sensitive to pain than boys (Myers, Riley, & Robinson, 2003; however, see Earp, 2016b for a discussion of different pain stereotypes in some non-Western contexts). Thus, one might have expected participants to rate the child as experiencing more pain when it was described as a girl rather than the other way around. However, the finding from Cohen et al. (2014) can plausibly be explained in light of equally common social norms concerning the appropriate pain response for boys when compared with girls. As Cohen et al. (2014) argue: "socialization based on gender may influence gender-based pain expression in which boys learn to display stoicism in response to pain whereas girls may engage in more expressive responses to encourage support." In addition, "both males and females have stereotypical beliefs regarding pain tolerance, such that males are perceived to be more tolerant of pain" (p. 93).

Thus, the findings reported by Cohen et al. (2014) could reflect salient sociocultural norms according to which it is less acceptable for boys to display overt pain behaviors than it is for girls (Bernardes et al., 2008; Robinson & Wise, 2003). Since the child in the video was clearly displaying pain (e.g., saying "Ow!"), participants in the Boy condition might have inferred—whether consciously or unconsciously—that he must *really* be in pain to behave in such a way. Whereas, in the Girl condition, participants might have inferred that less actual pain would be needed to elicit the observed behaviors.

Our proposed inferential model thus has the following structure: adults first notice a display of pain produced by a particular child. They then use various sources of information in their environment, in conjunction with relevant prior beliefs, to make a judgment about the trustworthiness of that display as an indicator of the underlying sensation of pain actually felt by the child. For example, they might judge that the child is overreacting (displaying more pain than is actually felt), underreacting (displaying less pain than is actually felt), or accurately reacting (displaying the same amount of pain as is actually felt). The adult then discounts, augments, or accepts the level of pain displayed in the child's behavior, and judges the felt level of pain accordingly.

Based on the research reviewed previously, it is reasonable to think that stereotyped gender-role expectations concerning the appropriate or at least typical pain response for boys as compared to girls will play a role in such judgments (Hoffmann & Tarzian, 2001). Specifically, when the male gender role is salient, the inferred level of pain sensation from a fixed display of pain should be relatively high (and vice versa with the female gender role). In line with this hypothesis, we sought to replicate, clarify, and extend the preliminary finding of Cohen et al. (2014) as well as to assess its generalizability in a larger and more demographically diverse population.

To do this, we experimentally manipulated the perceived gender of a target child while holding the display of pain constant across conditions, and measured the influence of this manipulation on adult assessments of the child's pain experience. We had two predictions: first, that when the child was described as a boy as opposed to a girl, "he" would be rated as experiencing more pain; and second, that this effect would be eliminated when controlling for participants' explicit gender stereotypes concerning the expression of pain.

Study 1

The purpose of this study was to determine whether or to what extent adult participant judgments of a target child's pain would differ depending on whether the target was described as a boy or a girl. As a first step, we conducted a direct replication of Cohen et al. (2014) to establish the reliability of the basic effect, albeit with a larger and more demographically diverse sample (see Earp and Trafimow, 2015 for a discussion of replication terminology and the relevant context). We contacted the lead author of Cohen et al. (2014) to ask for the original materials and for guidance on running the study in a way that would be maximally faithful to the original apart from being adapted to an online computer interface. The author was very helpful and also shared the materials freely; it was also affirmed that our final design was adequate for purposes of fair replication. Otherwise, the original research team was uninvolved with the data collection, analysis, and writing of the first draft of the report. The lead author of Cohen et al. (2014) was then invited to join as a co-author, contributing to subsequent versions of the manuscript. To minimize potential researcher degrees of freedom (Simmons, Nelson, & Simonsohn, 2011),

we pre-registered our design and analysis plan with aspredicted.org. The pre-registration form, along with all materials, data, and syntax for reproducing the reported analyses are available at the Open Science Framework (OSF) project folder associated with this paper, at https://osf.io/t73c4/.

Methods

Participants

This study was approved by the Yale University Research Ethics Committee (Human Subjects Committee Protocol ##2000021893). Five hundred and two U.S. participants were recruited via Amazon's Mechanical Turk (MTurk) and received $1 for their time. MTurk is a "marketplace for work that requires human intelligence" (www.mturk.com), which matches requesters and workers for short-term tasks. Requesters, including psychology researchers, pay a fee to Amazon to post self-contained jobs on the platform, such as an experiment or a survey. Research suggests that MTurk participants are generally more demographically diverse than traditional student samples, while typically being more attentive (Hauser & Schwarz, 2016) and no less reliable (Buhrmester, Kwang, & Gosling, 2011). The platform has become a standard research tool across the social sciences (Buhrmester, Talaifar, & Gosling, 2018).

To ensure high-quality data from this online sample, we pre-registered extremely strict exclusion criteria, based on multiple attention, comprehension, and other checks embedded throughout the survey. Thus, we erred on the side of caution, and recruited roughly twice as many participants as we needed according to a power analysis based on the effect size reported in Cohen et al. (2014) (i.e., 500 for a target sample of 260); this analysis was conducted using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007), with Cohen's $d = .31$, $\alpha = .05$, and power = .80.

Because this was an online replication of a study originally conducted in-person in a laboratory, we wanted to make sure that participants were fully attending to, understanding, and complying with the given instructions. Therefore, we included several attention and comprehension checks in addition to an audio check and a manipulation check (see Table I). Ultimately, 238 participants were excluded for failing one or more of these criteria (i.e., some participants are double or triple counted in the exclusions listed in Table I, having failed more than one). The final sample consisted of 264 participants (136 female, 128 male) ranging in age from 18 to 75 ($M = 38.05$, $SD = 11.21$). See Table II for complete demographic information. Participants were randomly assigned to one of two conditions, Boy ($n = 133$) or Girl ($n = 131$), reflecting described target gender.

Procedure

Participants completed an online survey with both between-subjects and within-subjects components. Depending on condition, they were told "You are about to watch a short video clip of a little girl [boy] receiving a fingerstick to test iron levels during her [his] doctor visit." They then saw a short clip of a 5-year-old child whose perceivable gender characteristics were deemed by Cohen et al. (2014) to be ambiguous between male and female in pre-testing—a topic to which we will return in the discussion. The clip was the same one used in Cohen et al. (2014). Please note that the child was consulted about and agreed to the use of the video for research purposes; the actual video is not included in the online materials to preserve the child's privacy. However, a protected link for viewing can be shared upon request.

In the clip, the child is undergoing a short medical procedure involving a finger-stick to draw blood. Participants were shown the same clip, regardless of whether they were in the Boy or Girl condition. They were instructed to pay close attention to how much pain the child was *sensing* (experiencing) and *displaying* (showing).[1] We defined *sensation of pain* as "a measure of how much pain a person actually experiences" and *display of pain* as "a measure of how much people express that they are in pain. This includes behavior such as crying, grimacing, or saying it hurts." Depending on condition, participants were then told "You just watched a video clip of Samantha [Samuel] receiving her [his] Pre-Kindergarten finger stick," and were then asked a range of questions about how much pain they thought the child in the clip was sensing and displaying. Participants were then asked to report their explicit beliefs regarding pain reactions in boys versus girls as well as the typicality of the target child in these respects, followed by demographic measures.

Measures

Pain Sensation and Pain Display. Participants were given two questions designed to capture their judgments about how much pain the child in the clip was sensing and displaying. For both questions, participants were given a sliding Visual Analogue Scale (VAS), a measure often used in pain research (Bourdel et al., 2015). The questions were presented in a fixed order as follows: (a) "How much pain did she [he] *experience* during the finger stick?" (*Pain Sensation*) and (b) "How much pain did she [he] *display* during the finger stick?" (*Pain Display*).

1 As a reviewer notes, the use of two questions—for pain sensation *and* pain display—could potentially predispose participants to observe the child or respond differently than they might in real life, where consciously distinguishing between these facets is presumably less likely to happen.

Table I. *Summary of Excluded Data*

Exclusion criterion met	Type of check	Excluded *N*	Included in
Failed to identify at least one definition or measure presented on previous screen(s) but met no further exclusion criteria	Attention check	63	Excluded from all pre-registered analyses. Reintroduced for exploratory robustness analysis
Failed to identify target child gender or hair color	Manipulation check	103	Not included in any analysis
Failed audio test (listen to a pre-recorded sentence and identify the correct sentence among options)	Equipment check	10	Not included in any analysis
Failed to spend at least 4 min completing the survey	Quality check	90	Not included in any analysis
Failed age or English fluency requirements (mark being age 18 or older and fluent in English)	Demographic check	2	Not included in any analysis

Note. Some participants failed two or more criteria and were thus counted double in the table.

Table II. *Participant Demographic Characteristics for Study 1*

Age	Frequency	Percentage	Race/Ethnicity	Frequency	Percentage
18–25	25	9.5	Black/African American	23	8.7
26–35	110	41.7	Asian	16	6.1
36–45	70	26.5	Hispanic/Latinx	12	4.5
46–55	31	11.7	Hawaiian/Pacific Islander	1	.4
56–65	23	8.7	White	206	78.0
66–75	4	1.5	Other	6	2.3
Missing	1	0.4			
Total	264	100	Total	264	100
Gender	**Frequency**	**Percentage**	**Parental Status**	**Frequency**	**Percentage**
Female	136	51.5	Children	126	47.7
Male	128	48.5	No Children	138	52.3
Total	264	100.0	Total	264	100.0

The male or female pronoun was used depending on condition. The VAS ranged from 0 to 100, with 0 labeled "No Pain" and 100 labeled "Severe Pain." It should be noted that Cohen et al. (2014) did not include the second measure (*Pain Display*) in their original study. In consultation with Cohen, however, we added this measure so that we would have a complete set of observational-judgment measures and explicit belief measures for both sensation and display (see below for further discussion).

Explicit Sensation and Explicit Display. Participants were given two questions designed to capture their explicit beliefs about differences in pain sensation and pain display between boys and girls in general. Consistent with Cohen et al. (2014), these questions were taken from the Gender Role Expectations of Pain Questionnaire (Robinson et al., 2001) but were modified to fit the VAS format. As in the Cohen et al. study, the questions were presented in a fixed order as follows: (a) "In general, compared with girls, boys' *sensation of pain* is …" (*Explicit Sensation*) and (b) "In general, compared with girls, boys' *display of pain* is …" (*Explicit Display*). These questions were the same across both conditions. Participants were given a sliding scale ranging from 0 to 100, with 0 labeled "Far Less," 50 labeled "The Same," and 100 labeled "Far Greater."

Target Typicality. Participants were next shown four measures assessing the typicality of the target child (i.e., how closely their sensation and display of pain aligned with what is typical for a boy or girl). These measures were included in the original study materials given to us by the lead author of Cohen et al., (2014), although no associated results were reported in the published paper. We decided to include the measures for purposes of exploratory analysis. The questions were presented in a fixed order as follows: (a) "Compared with the typical girl, the child in the video's *sensation of pain* was …," (b) "Compared with the typical boy, the child in the video's *sensation of pain* was …," (c) "Compared with the typical girl, the child in the video's *display of pain* was …," and (d) "Compared with the typical boy, the child in the video's *display of pain* was …" These questions were the same across both conditions. Participants were given a sliding scale ranging from 0 to 100, with 0 labeled "Far Less," 50 labeled "The Same," and 100 labeled "Far Greater."

Other questions—drawn from the original Cohen et al. (2014) materials—that were included but not

analyzed for the present report were: "How anxious was [s]he *before the finger stick?"*, "How anxious was [s]he *during the finger stick*?", "How anxious was [s]he *after the finger stick*?", and "How anxious were *you watching* this video?" The reason we did not analyze these data is because we had no *a priori* hypothesis concerning them and we wanted to run as few statistical tests as possible for purposes of replication so as not to inflate the chance of a Type 1 error for peripheral results.

Results

For the following analyses, the alpha level was set at .05 based on the criterion used by Cohen et al. (2014), with *p*-values falling below this threshold defined in advance as statistically significant. Note: "statistically significant" does not entail "clinically significant." The goal of the present study is to test new-sample effect replicability of an original finding (see LeBel et al., 2018 for a discussion of this terminology); assessing clinical or practical implications is a task for future research.

Pain Sensation

As pre-registered, a two (participant gender: male, female) by two (target gender: boy, girl) analysis of variance (ANOVA) was conducted to test for an interaction between participant gender and target gender on judgments of pain sensation. Neither the main effect of participant gender, $F(1, 260) = 1.34$, $p = .248$, $\eta_p^2 = .005$, nor its interaction with condition, $F(1, 260) = 2.61$, $p = .107$, $\eta_p^2 = .010$ was statistically significant; we therefore collapsed across participant gender and performed a *t*-test to see whether ratings of pain sensation differed between conditions.

Consistent with Cohen et al. (2014), ratings for pain sensation were significantly higher in the Boy condition ($M = 50.42$, $SD = 20.35$) than in the Girl condition ($M = 45.90$, $SD = 22.12$), $t(262) = -1.73$, $p = .043$ (one-tailed), $d = .21$, 95% confidence interval (CI) $[-\infty, -.20]$. In other words, participants rated the child as experiencing more pain when it was described as a boy as compared to a girl. The one-tailed *p*-value we report here is due to having pre-registered a directional hypothesis for the main finding, favoring a reduction in risk of committing a Type II versus Type I error for purposes of replication (see Boyle, 2018). Note that the absolute scores in our study were lower (i.e., shifted down the scale) compared with the original (Figure 1).

Before proceeding further, we performed a visual inspection of the data. Although as noted we did not find a moderating effect of participant gender on judgements of pain sensation, a look at the means for male and female participants as a function of condition strongly suggests that female participants drove

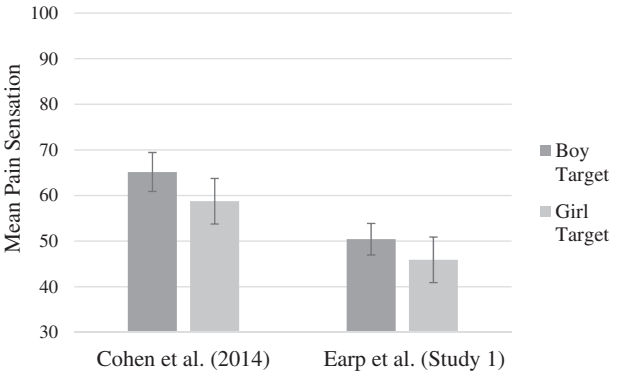

Figure 1. Mean pain sensation scores for Cohen et al. (2014) and Earp et al. (Study 1). Error bars are 95% CI. Cohen et al. (2014) reported higher pain scores for the Boy target ($M = 65.15$; $SD = 20.77$) than the Girl target ($M = 58.75$; $SD = 20.83$), $t(181) = 2.07$, $p = .04$, with no moderation by participant gender. Please note that the *y*-axis has been truncated for ease of interpretation. CI = confidence interval.

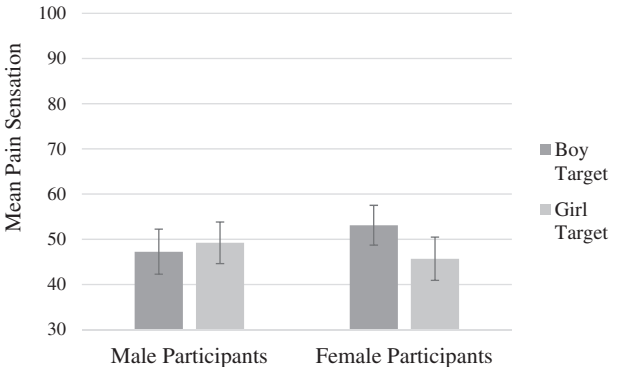

Figure 2. Mean pain sensation scores for Study 1: male and female participants as a function of condition. Error bars are 95% CI. The *y*-axis has been truncated for ease of interpretation. CI = confidence interval.

the main effect in our study (Figure 2). To assess this possibility, we conducted an exploratory robustness analysis which we had not pre-registered. Such an analysis (also sometimes called a sensitivity analysis), is especially desirable as a complement to confirmatory analyses conducted after a large number of participants have been excluded based on pre-registered criteria. This is to ensure that any resultant findings are robust against different sets of plausible exclusions (Schuwerk, Priewasser, Sodian, & Perner, 2018; Thabane et al., 2013). Prior research with online samples suggests that MTurk workers who fail one or more embedded attention checks nevertheless often are paying sufficient attention to the rest of the study that their judgments remain informative; systematically excluding their data thus results in only negligible improvement to reliability (Rouse, 2015). For purposes of our exploratory robustness/sensitivity analysis, therefore, we reintroduced participants who

had failed one or more of the multiple attention checks but not the manipulation checks—nor met any other exclusion criteria—in order to increase power (Table I). This left us with a sample of $N = 327$ participants (170 female, 156 male, 1 other) ranging in age from 18 to 75 ($M = 37.15$, $SD = 11.03$).

As before, a two (participant gender: male, female) by two (target gender: boy, girl) ANOVA was conducted on judgments of pain sensation. This time, with the larger sample, the main effect of condition, $F(1, 322) = 1.28$, $p = .258$, $\eta_p^2 = .004$ was not reliable, but there was a significant main effect of participant gender, $F(2, 322) = 3.70$, $p = .048$, $\eta_p^2 = .019$, which was qualified by an interaction between participant gender and condition, $F(1, 322) = 3.83$, $p = .051$, $\eta_p^2 = .012$. Simple effects tests revealed no effect of condition among the male participants: $t(154) = .57$, $p = .571$, $d = .09$, 95% CI [−4.91, 8.86]; however, among female participants, those in the Boy condition rated the child as experiencing significantly more pain ($M = 53.10$, $SD = 20.43$) than those in the Girl condition ($M = 45.69$, $SD = 22.57$), $t(168) = -2.25$, $p = .026$, $d = .34$, 95% CI [−13.92, −.89]. This suggests that the main finding from the smaller, full-exclusion data set (Boy: $M = 50.42$, $SD = 20.35$; Girl: $M = 45.90$, $SD = 22.12$) was indeed likely driven by female participants. We note that the larger effect size estimate from this higher-powered analysis, $d = .34$, is closer to, and in fact slightly larger than, that reported by Cohen et al. (2014), namely $d = .31$.

Pain Display

Returning now to the smaller, full-exclusion data set to continue with our pre-registered analyses, we conducted a two (participant gender: male, female) by two (target gender: boy, girl) ANOVA on judgments of pain display. There was no main effect of participant gender, $F(1, 260) = .01$, $p = .925$, $\eta_p^2 < .001$, nor an interaction between participant gender and condition, $F(1, 260) = .27$, $p = .607$, $\eta_p^2 = .001$. Collapsing across participant gender, ratings for pain display were similar for the Boy ($M = 71.64$, $SD = 19.88$) and Girl ($M = 69.02$, $SD = 19.38$) conditions, $t(262) = -1.082$, $p = .280$, $d = .13$, 95% CI [−7.38, 2.14]. Please note that results from the larger data set used for the robustness/sensitivity analysis were similar to the results reported here from the full-exclusion data set (all ps $> .284$).

Explicit Sensation

Continuing with the stricter, smaller data set and our pre-registered analyses, a two (participant gender: male, female) by two (target gender: boy, girl) ANOVA was conducted on explicit sensation beliefs. There was no main effect of participant gender, $F(1, 260) = 1.17$, $p = .281$, $\eta_p^2 = .004$, nor an interaction between participant gender and condition, $F(1, 260) = 1.68$, $p = .197$, $\eta_p^2 = .006$. We therefore collapsed across participant gender and ran a t-test comparing explicit sensation beliefs against the midpoint of the scale ("In general, compared with girls, boys' sensation of pain is . . ." 50 = The Same).

Ratings for explicit sensation beliefs did not differ from the midpoint of the scale ($M = 49.63$, $SD = 11.73$): $t(263) = -.52$ $p = .604$, $d = .04$, 95% CI [−1.80, 1.05]. In other words, participants did not express any explicit beliefs that girls and boys experience pain differently in general. This result differs from that of Cohen et al. (2014), whose participants did explicitly state that girls are more sensitive to pain than boys ($M = 44.30$, $SD = 15.78$), $t(182) = 4.89$, $p = .001$, $d = .36$.

Unexpectedly, there was a main effect of condition (target gender) on explicit sensation judgments, suggesting a priming effect on explicit, general attitudes about the pain experiences of boys compared with girls: $F(1, 260) = 5.47$, $p = .020$, $\eta_p^2 = .021$. Specifically, when the target was described as a girl, participants rated boys as in general experiencing less pain than girls ($M = 47.82$, $SD = 12.84$), whereas when the target was described as a boy, participants rated boys as in general experiencing more pain than girls ($M = 51.41$, $SD = 10.25$). To see whether either of these means differed from the midpoint of the scale, we analyzed each condition separately. Within the Girl condition, a one-sample t-test did not show a significant difference: $t(130) = -1.95$, $p = .054$, $d = .24$, 95% CI [−4.40, .04], however, the p-value is very close to the alpha threshold. Within the Boy condition, there was similarly no significant difference: $t(132) = 1.58$, $p = .116$, $d = .19$, 95% CI [−.35, 3.16]. Together, it seems that there was a small priming effect in the Girl condition, whereby participants in that condition rated boys in general as experiencing less pain than girls.

Explicit Display

As pre-registered, a two (participant gender: male, female) by two (target gender: boy, girl) ANOVA was conducted on explicit display beliefs. There was no main effect of participant gender, $F(1, 258) = 3.46$, $p = .064$, $\eta_p^2 = .013$, nor an interaction between participant gender and condition, $F(1, 258) = .004$, $p = .953$, $\eta_p^2 < .001$. We therefore collapsed across participant gender and ran a t-test against the midpoint of the scale ("In general, compared with girls, boys' display of pain is . . ." 50 = The Same).

Ratings for explicit display beliefs did differ from the midpoint of the scale ($M = 44.95$, $SD = 18.00$), with participants rating boys as in general displaying less pain than girls: $t(261) = -4.54$, $p < .001$, $d = .40$, 95% CI [−7.24, −2.86]. This result supports the

findings of Cohen et al. (2014), who also reported that participants expressed such explicit beliefs ($M = 41.38$, $SD = 18.21$), $t(180) = 6.37$, $p < .001$, $d = .47$. Cohen et al. (2014) reported that this effect was significantly differentiated by participant gender, with male participants ($N = 28$) reporting a greater difference between the display of pain in boys and girls ($M = 33.33$, $SD = 19.73$) than female participants ($N = 124$, $M = 42.75$, $SD = 17.67$), $t(178) = 2.51$, $p = .013$, $d = .50$, 95% CI [1.95, 16.89]. Although our p-value for participant gender was not less than .05, it was near this threshold at .064. If we run the same t-test as Cohen et al. (2014) on our own data, we find a similar pattern of results, albeit with a smaller effect size. That is, male participants ($N = 128$) reported a greater difference between the display of pain in boys and girls ($M = 42.63$, $SD = 16.98$) than did female participants ($N = 134$, $M = 47.16$, $SD = 18.72$), $t(260) = -2.05$, $p = .041$, $d = .25$, 95% CI [.24, 8.82].

Target Typicality

Condition: Girl. Starting with sensation of pain, there was no effect of participant gender on ratings of target similarity to the typical girl ($p = .530$) or the typical boy ($p = .545$). Compared with the typical girl, the child in the video was rated as experiencing a similar amount of pain ($M = 50.96$, $SD = 11.47$): $t(130) = .96$, $p = .339$, $d = .12$, 95% CI [−1.02, 2.94]. Likewise compared with the typical boy: ($M = 51.43$, $SD = 11.65$): $t(130) = 1.40$, $p = .163$, $d = .17$, 95% CI [−.59, 3.44]. In other words, participants judged "Samantha" as experiencing pain in a manner that is no different from what is typical for a girl (or a boy).

For display of pain, there was also no effect of participant gender on ratings of target similarity to the typical girl ($p = .876$) or the typical boy ($p = .366$). Compared with the typical girl, the child in the video was rated as displaying a similar amount of pain ($M = 50.75$, $SD = 14.39$): $t(130) = .60$, $p = .553$, $d = .07$, 95% CI [−1.74, 3.24]. However, compared with the typical boy, the child was rated as displaying *more* pain ($M = 54.27$, $SD = 17.10$): $t(130) = 2.86$, $p = .005$, $d = .35$, 95% CI [1.32, 7.23]. This is consistent with our finding for explicit display beliefs, wherein participants rated girls in general as displaying more pain than boys. Taken together, participants rated "Samantha" as typical for a girl, both in terms of experience and display of pain.

Condition: Boy. Starting with sensation of pain, there was no effect of participant gender on ratings of target similarity to the typical girl ($p = .668$) or the typical boy ($p = .242$). Compared with the typical girl, the child in the video was rated as experiencing a similar amount of pain ($M = 50.99$, $SD = 9.40$): $t(132) = 1.21$, $p = .229$, $d = .15$, 95% CI [−.63, 2.60]. However, compared with the typical boy, the child in the video was rated as experiencing *more* pain ($M = 53.02$, $SD = 11.16$): $t(132) = 3.11$, $p = .002$, $d = .38$, 95% CI [1.10, 4.93]. Thus, even when participants were explicitly told that the child in the video was a boy, they rated "Samuel's" experience of pain as being greater than that of a typical boy, and similar to that of a typical girl.

For display of pain, there was again no effect of participant gender on ratings of target similarity to the typical girl ($p = .069$) or the typical boy ($p = .158$). Compared with the typical girl, the child in the video was rated as displaying a similar amount of pain ($M = 51.26$, $SD = 10.81$): $t(132) = 1.35$, $p = .180$, $d = .16$, 95% CI [−.59, 3.12]. However, compared with the typical boy, the child was rated as displaying *more* pain ($M = 53.43$, $SD = 12.46$): $t(131) = 3.16$, $p = .002$, $d = .39$, 95% CI [1.29, 5.58]. Thus, "Samuel" was *not* seen as being typical for a boy, either in terms of sensation or display of pain, both of which were rated as higher than what is typical for boys. Taken together with the results from the Girl condition, we find that, regardless of condition, the child in the video was rated as both sensing and displaying pain in a manner that is typical for a girl but greater than what is typical for a boy.

Controlling for Explicit Beliefs

To test the hypothesis that explicit beliefs or stereotypes concerning male versus female child pain display behaviors could account for the main finding of a difference in pain sensation scores between conditions, we controlled for such beliefs and re-ran the corresponding analysis (based on Martin & Ruble, 2010). We reasoned as follows: if the belief that boys tend to display less pain than girls is what is driving the relevant inferential process—that is, that this particular "boy" must *really* be in pain—then controlling for that belief should make the between-subjects difference in pain sensation ratings diminish or disappear. Indeed, this is what we find: when controlling for explicit gender stereotypes, the implicit effect of condition on inferred sensation of pain was no longer statistically significant: $F(1, 257) = 1.65$, $p = .200$, $\eta_p^2 = .006$.[2]

Discussion

In this direct replication and extension of Cohen et al. (2014), we used highly similar methods and materials and found evidence that is supportive of their main published finding, albeit in a larger and more

2 To be able to report interpretable 95% CIs for this analysis, we provide here the results of the same ANOVA on pain sensation (fixed factor: target gender, covariates: explicit display and sensation), conducted as a regression: $B = -3.45$, $SE = 2.65$, 95% CI [−8.67, 1.77], $t(258) = -1.30$, $p = 0.200$.

demographically diverse sample. In Cohen et al. (2014), when a target child was described as a boy, undergraduate participants rated the child as experiencing more pain than when the child was described as a girl. In our study, we found the same effect among US-based MTurk workers, but it was smaller in size and only statistically significant using a one-tailed *t*-test. However, an exploratory robustness analysis with fewer exclusions and increased power resulted in a larger effect size estimate comparable to that of Cohen et al. (2014), as well as a statistically significant *p*-value using a two-tailed *t*-test.

Notably, we found that this effect was driven by female participants. This finding is in some tension with that reported by Moon et al. (2008), who found that fathers, but not mothers, rated the pain of their sons higher in a cold pressor task, as well as Rosenbloom et al. (2011), who found no difference between fathers and mothers in ratings of their children's pain. However, the present study did not assess parental judgments of their own children's pain, but rather adult ratings of an unrelated child's pain. The results may therefore not be directly comparable.

Nevertheless, the finding that female participants drove our main effect is worth considering. As noted previously, the original study by Cohen et al. (2014) had a large number of female participants compared with male participants ($n = 123$ vs. $n = 28$). This was due to the authors' sampling of psychology and nursing students, who are disproportionately female. Thus, the possibility is raised that female participants were primarily responsible for their main finding as well, as the authors acknowledge (see Cohen et al., 2014, p. 93). Why it is that female, but not male, participants in our study rated "Samuel" as experiencing more pain than "Samantha" is not immediately clear, but this should be kept in mind and explored in future research.

An unexpected finding was that the child in the video was consistently rated as being more typical for a girl than a boy, in terms of both sensation and display of pain. As noted, we used the same video clip for our experimental stimulus as did Cohen et al. (2014), who through a pilot study had concluded that the target child's features and clothing were such that one could plausibly believe the child to be either a boy or a girl. The pilot study was informal and involved showing the video to students and asking them to guess the gender of the child. In reality, the child was a girl. As Cohen et al. (2014, p. 91) state: "The child was dressed in gender neutral clothing, consisting of a red tee-shirt and sports-style shorts. The child's hair partially covered her face, which made determining her gender difficult."

Yet there is a difference between a child being perceived as equally typical in appearance for either a boy or girl, and having an appearance (or behavior) that is simply sufficiently ambiguous that the child could plausibly be regarded as male or female depending on framing. If, in the absence any gendered framing, participants would overwhelmingly have seen the child as a girl—as she is in real life—this would complicate our interpretation of the findings.

This is for the following reason. Previous research suggests that children are socialized into distinct male or female gender roles starting at a very young age (Martin & Ruble, 2010). Such socialization includes learning the "appropriate" behavioral displays in response to pain for one's assigned gender category (e.g., "boys shouldn't cry"). Given that the child in the video was an actual 5-year-old girl, it is almost certain that she will have been socialized to some extent, thus raising questions about sociocultural factors which may have influenced her learned behavioral reactions to pain (Cohen et al., 2014, p. 93). In other words, it is possible that participants, rather than exhibiting a biased judgment grounded in unreflective gender stereotypes, were making a reasonable gender-based inference about the likely relationship between pain sensation and pain display in the video they watched, given real-life socialization pressures that tend to affect boys and girls differently.

On the other hand, if participants would have been roughly as likely to believe that the child was a boy or girl in the absence of gendered framing, their tendency to rate the child as more "typical" for a girl even in the Boy condition may indeed have reflected such gendered stereotypes. In other words, the overt display of pain they observed might have conflicted with the stoicism characteristically expected of boys, appearing more in line with the relative emotivism characteristically expected of girls. This, then, could lead participants to judge that this particular boy must *really* be in pain, as we hypothesized. To look into this issue, we conducted a short follow-up study with a new sample of MTurk workers.

Study 2

The purpose of Study 2 was to determine whether, without any gendered framing, participants would be far more likely judge the child in the stimulus video to be a girl as compared to a boy or vice versa (low gender ambiguity), or roughly equally likely to judge the child to be a girl or a boy (high gender ambiguity). As this study was exploratory, we did not pre-register a numerical cutoff for high or low ambiguity, nor did we run inferential statistics. Instead, we provide simple descriptives to aid interpretation.

Methods

Participants

Based on available funding, we recruited 117 new MTurk participants (42 female, 75 male), none of

whom had been involved in the previous study, and paid them 20 cents each for their time. Two were excluded for indicating that their age was below 18 and 5 were excluded for failing an embedded audio check, leaving a final sample of $N = 110$ (41 female, 69 male) ranging in age from 20 to 70 years ($M = 34.33$ $SD = 11.60$). Of these, 64.5% identified as White, 15.5% as Asian, 8.2% as Black/African American, 8.2% as Hispanic/Latinx, 1.8% American Indian/Alaska Native, 0.9% as Hawaiian/Pacific Islander, and 0.9% as Other.

Procedure

After an audio check to confirm that their sound was working, participants were told "You are about to watch a short video clip of a child receiving a fingerstick to test iron levels during a doctor visit. Please watch the video closely and think about how much pain you think the child is experiencing during the procedure." Participants were then shown the same video as in the original study. After watching the video, participants were asked, "What do you think the sex of the child in the video was?" They were given the options "Male" and "Female" with the order of presentation randomized for each participant. Participants were also asked to guess the age and ethnicity of the child as filler questions.

Results

We found that, absent any gendered framing, 58.2% of participants correctly judged that the child was female, while 41.8% thought the child was male. The percentages were very similar for male and female participants: 58.0% versus 58.5% correct judgments, respectively. Thus, although the distribution of judgments was not exactly evenly split between male and female, the child's appearance seems to have been characterized by a high degree of gender ambiguity.

Discussion

In line with our prediction, this finding suggests that the asymmetrical judgments concerning target typicality observed in the previous study may indeed reflect gendered stereotypes. Specifically: a "boy" exhibiting overt discomfort in response to a painful stimulus was judged to both feel and express pain in a manner that is more typical for a girl than a boy. Yet in the absence of any gendered framing, that same child was roughly equally likely to be perceived as a boy or girl in a new sample of naïve participants.

General Discussion

In this research, we found that adult participants rated a child undergoing a medical procedure as experiencing more pain when the child was described as a boy as opposed to a girl—despite identical behavior and circumstances—and that this effect was eliminated by controlling for explicit gender stereotypes concerning the pain responses of boys and girls. We therefore replicated the findings of Cohen et al. (2014), while providing a plausible explanation for their results. In addition, since we relied on a larger and more demographically diverse sample than those authors, we have shown for the first time that the effect can generalize beyond a university student population.

That being said, the participant observers in our study were neither parents of the target child, nor—to our knowledge—medical professionals, but rather unrelated virtual bystanders. We therefore cannot speak to the applicability of our findings to judgments made in the context of a parent-child relationship or in a healthcare consultation. Further studies should explore these real-life contexts. A further limitation of our study concerns its reliance on a single video, identical to that used in Cohen et al. (2014). Although this was necessary for purposes of direct replication, it does raise the possibility of a stimulus-specific effect. As noted previously, the face of the child in the video was partially covered by her hair. This likely increased the gender ambiguity of the stimulus; yet facial expressions of pain are often an important source of information for observer judgments of child pain. This should be kept in mind as a potential limitation of the stimulus used in this study.

Another potential concern with the stimulus is that the child in the video was rated as more typical for a girl than a boy in terms of pain response regardless of condition. This complicated our interpretation of the results of Study 1 for the reasons described previously. However, in Study 2, we sought to address this concern, finding that only 58.2% of participants correctly judged that the child was female when no gendered framing was employed. This suggests that the relevant gender cues—from clothing, behavior, and so on—were highly ambiguous as intended. Future studies should use multiple videos including a mix of both male and female children with tightly controlled gender cues between conditions. Other characteristics such as the child's age and race or ethnicity should also be varied, so that the boundary conditions for the observed effect can be determined.

In conclusion, fair and appropriate treatment of pediatric pain requires judgments about children's internal mental states that are as free from distorting biases as possible. We have provided evidence that the perceived gender of a child—holding all else constant in an experimental design—can make a difference to how their pain is interpreted by observers. Attempting to understand the specific factors that play into such differences across a range of demographic factors, and with respect to various types of painful experiences, should be a priority for researchers working in this

area. For healthcare providers, reliable information about potential biases in judgments concerning the private pain experiences of their patients—especially children—will be an important step in improving diagnosis and treatment going forward.

Funding

Funding for this research was provided by the Yale University Psychology Department.

Conflicts of interest: None declared.

References

Barrett, L. F., & Bliss-Moreau, E. (2009). She's emotional. He's having a bad day: Attributional explanations for emotion stereotypes. *Emotion (Washington, D.C.)*, *9*, 649–658. doi:10.1037/a0016821

Bernardes, S. F., Keogh, E., & Lima, M. L. (2008). Bridging the gap between pain and gender research: A selective literature review. *European Journal of Pain (London, England)*, *12*, 427–440. doi:10.1016/j.ejpain.2007.08.007

Bernardes, S. F., & Lima, M. L. (2011). On the contextual nature of sex-related biases in pain judgments: The effects of pain duration, patient's distress and judge's sex. *European Journal of Pain*, *15*, 950–957. doi:10.1016/j.ejpain.2011.03.010

Boerner, K. E., Birnie, K. A., Caes, L., Schinkel, M., & Chambers, C.T. (2014). Sex differences in experimental pain among healthy children: A systematic review and meta-analysis. *Pain*, *155*, 983–993. doi:10.1016/j.pain.2014.01.031

Bourdel, N., Alves, J., Pickering, G., Ramilo, I., Roman, H., & Canis, M. (2015). Systematic review of endometriosis pain assessment: How to choose a scale? *Human Reproduction Update*, *21*, 136–152. doi:10.1093/humupd/dmu046

Boyle, G. J. (2018). Proving a negative? Methodological, statistical, and psychometric flaws in Ullmann et al. (2017) PTSD study. *Journal of Clinical and Translational Research*, *3*(S2), 375–381.

Buhrmester, M., Kwang, T., & Gosling, S.D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*, 3–5. doi:10.1177/1745691610393980

Buhrmester, M., Talaifar, S., & Gosling, S.D. (2018). An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science*, *13*, 149–154.

Cohen, L. L., Cobb, J., & Martin, S. R. (2014). Gender biases in adult ratings of pediatric pain. *Children's Health Care*, *43*, 87–95. doi:10.1080/02739615.2014.849918

Cohen, L. L., Lemanek, K., Blount, R. L., Dahlquist, L. M., Lim, C. S., Palermo, T. M., … Weiss, K. E. (2008). Evidence-based assessment of pediatric pain. *Journal of Pediatric Psychology*, *33*, 939–955. doi:10.1093/jpepsy/jsm103

Craig, K. D. (2018). Toward the social communication model of pain. In T. Vervoort, K. Karos, Z. Trost, & K. M. Prkachin (Eds.), *Social and interpersonal dynamics in pain* (pp. 23–41). Cham: Springer International Publishing. doi:10.1007/978-3-319-78340-6_20

Earp, B. D. (in press). The child's right to bodily integrity. In D. Edmonds (Ed.), *Ethics and the Contemporary World*. Abingdon and New York: Routledge. Retrieved from https://www.academia.edu/37138614/The_childs_right_to_bodily_integrity

Earp, B. D. (2016a, July 2). In praise of ambivalence—"young" feminism, gender identity and free speech. Retrieved from http://blog.practicalethics.ox.ac.uk/2016/07/in-praise-of-ambivalence-young-feminism-gender-identity-and-free-speech/ Retrieved 13 September 2018.

Earp, B. D. (2016b). Between moral relativism and moral hypocrisy: Reframing the debate on "FGM." *Kennedy Institute of Ethics Journal*, *26*, 105–144. doi:10.1353/ken.2016.0009

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 621–611. doi:10.3389/fpsyg.2015.00621

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. doi:10.3758/BF03193146

Freeman, J. B., Rule, N. O., Adams, R. B., & Ambady, N. (2010). The neural basis of categorical face perception: Graded representations of face gender in fusiform and orbitofrontal cortices. *Cerebral Cortex*, *20*, 1314–1322. doi:10.1093/cercor/bhp195

Harrison, J., Grant, J., & Herman, J. L. (2012). A gender not listed here: Genderqueers, gender rebels, and otherwise in the National Transgender Discrimination Survey. *LGBTQ Policy Journal at the Harvard Kennedy School*, *2*, 13–24.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*, 400–407. doi:10.3758/s13428-015-0578-z

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. Tory Higgins and Arie W. Kruglansk (Eds.), *Social psychology: Handbook of basic principles* (pp. 133–168). New York, NY: Guilford Press.

Hirsh, A. T., Alqudah, A. F., Stutts, L. A., & Robinson, M. E. (2008). Virtual human technology: Capturing sex, race, and age influences in individual pain decision policies. *Pain*, *140*, 231–238. doi:10.1016/j.pain.2008.09.010

Hoffmann, D. E., & Tarzian, A. J. (2001). The girl who cried pain: A bias against women in the treatment of pain. *Journal of Law, Medicine & Ethics*, *28*, 13–27.

Hutson-Comeaux, S. L., & Kelly, J. R. (2002). Gender stereotypes of emotional reactions: How we judge an emotion as valid. *Sex Roles*, *47*, 1–10. doi:10.1023/A:1020657301981

Keogh, E. (2018). Sex and gender as social-contextual factors in pain. In T. Vervoort, K. Karos, Z. Trost, & K. M. Prkachin (Eds.), *Social and interpersonal dynamics in pain* (pp. 433–453). Cham: Springer International Publishing. doi:10.1007/978-3-319-78340-6_20

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389–402.

Martin, C. L., & Ruble, D. N. (2010). Patterns of gender development. *Annual Review of Psychology*, *61*, 353–381. doi:10.1146/annurev.psych.093008.100511

Moon, E. C., Chambers, C. T., Larochette, A.-C., Hayton, K., Craig, K. D., & McGrath, P. J. (2008). Sex differences in parent and child pain ratings during an experimental child pain task. *Pain Research & Management*, *13*, 225–230.

Musey, P. I., Linnstaedt, S. D., Platts-Mills, T. F., Miner, J. R., Bortsov, A. V., Safdar, B., … McLean, S. A. (2014). Gender differences in acute and chronic pain in the emergency department: Results of the 2014 Academic Emergency Medicine Consensus Conference pain section. *Academic Emergency Medicine*, *21*, 1421–1430. doi:10.1111/acem.12529

Myers, C. D., Riley, J. L., & Robinson, M. E. (2003). Psychosocial contributions to sex-correlated differences in pain. *The Clinical Journal of Pain*, *19*, 225–232.

Racine, M., Tousignant-Laflamme, Y., Kloda, L. A., Dion, D., Dupuis, G., & Choinière, M. (2012). A systematic literature review of 10years of research on sex/gender and pain perception—Part 2: Do biopsychosocial factors alter pain sensitivity differently in women and men? *Pain*, *153*, 619–635. doi:10.1016/j.pain.2011.11.026

Richards, C., Bouman, W. P., Seal, L., Barker, M. J., Nieder, T. O., & T'Sjoen, G. (2016). Non-binary or genderqueer genders. *International Review of Psychiatry*, *28*, 95–102. doi:10.3109/09540261.2015.1106446

Robinson, M. E., Riley, J. L., Myers, C. D., Papas, R. K., Wise, E. A., Waxenberg, L. B., & Fillingim, R. B. (2001). Gender role expectations of pain: Relationship to sex differences in pain. *The Journal of Pain*, *2*, 251–257. doi:10.1054/jpai.2001.24551

Robinson, M. E., & Wise, E. A. (2003). Gender bias in the observation of experimental pain. *Pain*, *104*, 259–264. doi:10.1016/S0304-3959(03)00014-9

Rosenbloom, E., Goldman, M., Konki, N., Edelman, S., Baram, W., & Kozer, E. (2011). Parental sex and age: Their effect on pain assessment of young children. *Pediatric Emergency Care*, *27*, 266. doi:10.1097/PEC.0b013e3182131438

Rouse, S. V. (2015). A reliability analysis of Mechanical Turk data. *Computers in Human Behavior*, *43*, 304–307. doi:10.1016/j.chb.2014.11.004

Schäfer, G., Prkachin, K. M., Kaseweter, K. A., & Williams, A. C. (2016). Health care providers' judgments in chronic pain: The influence of gender and trustworthiness. *Pain*, *157*, 1618. doi:10.1097/j.pain.0000000000000536

Schuwerk, T., Priewasser, B., Sodian, B., & Perner, J. (2018). The robustness and generalizability of findings on spontaneous false belief sensitivity: A replication attempt. *Royal Society Open Science*, *5*, 1–16.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366. doi:10.1177/0956797611417632

Sorge, R. E., & Strath, L. J. (2018). Sex differences in pain responses. *Current Opinion in Physiology*, *6*, 75–81. doi:10.1016/j.cophys.2018.05.006

Steinfeld, R., & Earp, B.D. (2017, May 15). How different are female, male and intersex genital cutting? Retrieved from http://theconversation.com/how-different-are-female-male-and-intersex-genital-cutting-77569 Retrieved 16 May 2017.

Thabane, L., Mbuagbaw, L., Zhang, S., Samaan, Z., Marcucci, M., Ye, C., … Goldsmith, C.H. (2013). A tutorial on sensitivity analyses in clinical trials: The what, why, when and how. *BMC Medical Research Methodology*, *13*, 1–12. doi:10.1186/1471-2288-13-92

Vervoort, T., Huguet, A., Verhoeven, K., & Goubert, L. (2011). Mothers' and fathers' responses to their child's pain moderate the relationship between the child's pain catastrophizing and disability. *Pain*, *152*, 786–793. doi:10.1016/j.pain.2010.12.010

Wise, E. A., Price, D. D., Myers, C. D., Heft, M. W., & Robinson, M. E. (2002). Gender role expectations of pain: Relationship to experimental pain perception. *Pain*, *96*, 335–342. doi:10.1016/S0304-3959(01)00473-0