Sex differences in humor production ability: A meta-analysis

By: Gil Greengross, Paul J. Silvia, and Emily C. Nusbaum

Greengross, G., Silvia, P. J., & Nusbaum, E. C. (2020). Sex differences in humor production ability: A meta-analysis. *Journal of Research in Personality, 84*, 103886. https://doi.org/10.1016/j.jrp.2019.103886

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

*****© 2019 Elsevier Inc. Reprinted with permission. This version of the document is not the version of record. *****

Abstract:

We offer the first systematic quantitative meta-analysis on sex differences in humor production ability. We included studies where participants created humor output that was assessed for funniness by independent raters. Our meta-analysis includes 36 effect sizes from 28 studies published between 1976 and 2018 (N = 5057, 67% women). Twenty of the 36 effect sizes, accounting for 61% of the participants, were not previously published. Results based on random-effects model revealed that men's humor output was rated as funnier than women's, with a combined effect size d = 0.321. Results were robust across various moderators and study characteristics, and multiple tests indicated that publication bias is unlikely. Both evolutionary and cultural explanations were considered and discussed.

Keywords: Humor | Humor production ability | Sex differences | Evolutionary psychology | Meta-analysis

Article:

1. Introduction

There is an ongoing debate about whether men and women differ in their cognitive abilities, how big the differences are, and how to explain them, if they exist (Halpern et al., 2007, Halpern, 2011, Hyde, 2005, Hyde, 2014, Lindberg et al., 2010, Lippa et al., 2010, Spelke, 2005, Voyer et al., 2007, Voyer et al., 1995, Voyer et al., 2017, Zell et al., 2015). Humor production ability (HPA)—a cognitive trait defined as the ability to produce funny remarks, create funny ideas, and make others laugh—is one such domain (Greengross and Miller, 2011, Greengross, 2014, Hooper et al., 2016, Lampert and Ervin-Tripp, 1998, Martin, 2014). Social stereotypes about sex differences in humor—particularly the stereotype that "women are not funny"—are culturally pervasive (Hitchens, 2007, Shlesinger, 2017). To date, no systematic review has evaluated the evidence for whether men and women differ in their humor production ability. After reviewing different schools of thought that seek to explain the role of biological sex in humor production, we meta-analytically synthesize the literature that has accumulated that can inform this question.

1.1. Considering sex differences in humor production ability

1.1.1. Evolutionary explanations

The universality of humor, its early developmental onset, and the fact that humans are not the only species that smiles and laughs have led researchers to suggest that humor has an evolutionary basis (Alexander, 1986, Chafe, 1987, Davila-Ross et al., 2009, Gamble, 2001, Hurley et al., 2011, Miller, 2000a, Preuschoft and Van-Hooff, 1997, Ramachandran, 1998, Viana, 2017, Weisfeld, 1993). The most relevant evolutionary theory pertaining to the possibility of sex differences in HPA is the mental fitness indicator theory, an extension of sexual selection theory (Darwin, 1871, Greengross and Miller, 2011, Howrigan and MacDonald, 2008, Miller, 2000a, Miller, 2000b).

According to sexual selection theory, males' and females' distinct behaviors and preferences are shaped due to asymmetrical parental investment (Trivers, 1972). In sexually reproducing species, the sex that bears the higher costs of reproduction is the choosier one, in most cases the females. Miller, 2000a, Miller, 2000b proposed that various cognitive capacities, such as language, arts, sports, and humor, evolved through mutual mate choice to advertise mate quality. These hard-to-fake mental fitness indicators serve to promulgate one's cognitive prowess, and are honest signals of intelligence that underline an individual's genetic quality. Humor is hypothesized to be one such fitness indicator, and HPA is positively correlated with various intelligence measures, most strongly with verbal aptitude (Christensen et al., 2018, Greengross and Miller, 2011, Howrigan and MacDonald, 2008, Kellner and Benedek, 2017). The mental fitness indicator theory proposed by Miller, 2000a, Miller, 2000b predicts that since women are choosier than men, men will be more motivated to advertise HPA in an effort to attract women, while women will put more effort in selecting mates based on men's ability to produce and showcase high levels of HPA. Why should this lead to higher levels of HPA among men? The answer lies in selection pressures involving a stronger male-male competition (intra-sexual selection), which drives men to ever improve their HPA in an effort to be funnier than their rivals in an attempt to attract women. Thus, we expect men to have a better HPA than women, as there is weaker evolutionary pressure for women to use humor to attract mates (Miller, 2000b).

Several studies support the various predictions stemming from the evolutionary explanation for viewing HPA as a sexually selected trait. Across cultures, sense of humor is found to be a more desirable trait in a mate for women choosing a mate, than for men (Buss, 1988, Feingold, 1992, Goodwin, 1990, Lippa, 2007, Sprecher and Regan, 2002, Todosijević et al., 2003, Toro-Morn and Sprecher, 2003). However, one study found no difference (McGee & Shevlin, 2009), and another found the opposite trend (i.e., that men view women as more attractive and more suitable as mates, when the women portrayed a great sense of humor, but not the opposite) (Antonovici & Turliuc, 2017). The apparent contradictory results could be explained by the lack of clarity in the meaning of the term 'sense of humor'. Saying that someone has a great sense of humor could mean that the person exhibits a high level of HPA, or that he or she enjoys humor or laughs often and easily. The lack of distinction leaves participants to have their own interpretation of the term and to mixed results. To resolve the issue, Bressler, Martin, and Balshine (2006) specifically tested whether women prefer men who display high HPA, and men

prefer women who appreciate their humor. The study found that although both men and women valued a good sense of humor in their respective partners, women showed a preference for a man with great HPA over a man that appreciated their humor production, while men preferred a woman that would appreciate their humor over a woman that would make them laugh. These results support the notion that when men and women talk about wanting a partner with a great sense of humor, they mean vastly different things. Men want a humor appreciator, while women want a humor producer.

Other studies from more ecologically valid situations, such as personal ads in newspapers and online dating sites, where people have low incentive to lie about their true preferences, show that women seek a mate who portrays humor ability twice as much as men do, and that men are more likely to declare how funny they are, or attempt humor, compared to women (Smith et al., 1990, Wilbur and Campbell, 2011). Women, on the other hand, express a desire for mates that will make them laugh, much more than men do, corroborating the prediction that men try to advertise their humor ability, and women are the appreciators of humor. Lastly, research shows a direct link between HPA and mating success. Adding humor to personal ads made men more attractive to women, but had little effect on women's attractiveness (Wilbur & Campbell, 2011). In another study, individuals who were rated high for HPA reported higher mating success as measured by number of sexual partners, age of first intercourse, and more sexual encounters, compared to individuals with low HPA (results were true for both sexes) (Greengross & Miller, 2011). However, women who have humorous partners did report having more and stronger vaginal orgasms, compared to women who have less funny partners, while men's sexual satisfaction was not related to women's HPA (Gallup, Ampel, Wedberg, & Pogosjan, 2014). These results highlight the significance of partner's high HPA on women, something that may contribute to higher reproductive success.

1.1.2. Social factors and humor production ability

Many social and cultural factors influence the way men and women create humor and how it is perceived. The notion that men are funnier than women is a widely-held stereotype and a cultural trope (e.g., Hitchens, 2007, Shlesinger, 2017). For example, when asked to describe an individual with a great sense of humor, or to name which sex is funnier, both men and women are much more likely to describe or choose a man (Crawford and Gressley, 1991, Nevo et al., 2001). In one study, 94% of the men and 89% of the women agreed to the stereotype that men are funnier than women (Mickes, Walker, Parris, Mankoff, & Christenfeld, 2012). In another study with both sexes surveyed, 62% of the participants believed that men have a greater HPA than women, 34% thought that men and women are equally funny, and only 4% viewed women as the funnier sex (Hooper et al., 2016). In addition, both men and women are more likely to attribute funny captions to male writers, and non-funny cartoons to women, even where the identity of the humor producer is concealed (Mickes et al., 2012). Such a stereotype may suppress women's willingness, and hinder their ability to create humor, ultimately putting them at a disadvantage compared to men.

Social role theory may be the best cultural framework in which observed sex differences are understood, and offers an alternative to the evolutionary explanation (Archer, 1996). According to the theory, sex differences emerge as part of the historical demarcation of men's and women's

roles within a society, which place them in unequal positions. In many societies, men have higher status and hold more power than women, while controlling the majority of the resources (Eagly & Wood, 1999). This power asymmetry leads the more powerful men to more masculine and dominant behaviors, while less powerful women exhibit behaviors that are subordinate and passive. Societal expectations to fit into sex-specific roles put pressure on both men and women to acquire the skills and adopt behaviors that will conform to their role requirements. If the stereotype that men have higher HPA than women is pervasive, cultural practices will work to sustain such a notion, and both men and women might try to fit into that expectation. Such practices were common throughout history and may still persist today. For example, for many years women tended to be the objects of jokes, often disparaging and sexist in nature, but rarely the subject producing the humor (Kotthoff, 2006). Specifically, women were prevented from using humor in the public sphere, not allowed to tell jokes and perform comedy routines, and confined to tell jokes only in private, while men were free to exhibit their humor in any form and platform they wished. These expectations, especially if indoctrinated from early ages, may contribute to observed sex differences in HPA.

The mechanisms in which such expectations transfer into behavior are often referred to as cultural scripts (Goddard & Wierzbicka, 2004). Cultural scripts include norms, values and practices that serve as collective guides to people on how to behave. Cultural scripts are highly influenced by gender stereotypes and could apply to various uses of humor. A specific script, the traditional courtship script, may come into play when using humor to attract mates. In the traditional courtship script, men are expected to be more active and take the initiative, while women are assumed to have a more passive role, being the recipients of men's romantic invitations (Eaton & Rose, 2011). In relation to humor, research suggests that men use humor to attract women, while women serve as appreciators by evaluating men's humor, thus fulfilling traditional gender role expectations (Wilbur & Campbell, 2011). How men and women view HPA in the traditional courtship script may be connected to traditional views of masculinity and femininity, with HPA being a masculine trait and humor appreciation viewed as feminine. Ross and Hall (in press) found that for both sexes, producing high quality humor in a courtship setting was associated with trait masculinity, with high masculine participants reporting using humor more to attract mates, compared to less masculine people of the same sex. In addition, these men also believed their HPA was better than same sex peers. Moreover, women adhering to traditional courtship behaviors, such as not making the first move when initiating a relationship, were less likely to use humor to attract men. Thus, the use of humor in courtship conforms to traditional gender roles of men as the pursuer and women the appreciator, with HPA to attract mates perceived as a masculine trait. The study illustrates the possible influences of sexual courtship scripts on how HPA is perceived, regardless of its real quality, and more generally, how cultural norms may affect how both men and women use humor.

1.2. Assessing humor production ability

Understanding the empirical literature on sex or gender, and humor production requires understanding how HPA is commonly assessed. As illustrated in Fig. 1, the measurement of humor creation ability typically consists of four steps, two for the creation of the humor, and two for the evaluation of the humor. First, a participant attempts to produce a humorous response to a non-funny stimulus or prompt provided by the researchers. Second, judges rate the participants'

responses, and their ratings for each participant are summarized to create the individual HPA score.

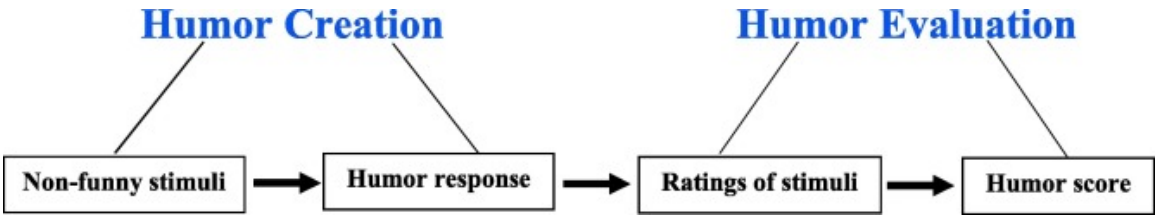

Figure 1. Measuring humor production ability. Participants are asked to generate humor in response to a non-funny stimulus. The responses are later evaluated by independent judges, and an overall score of humor production ability is calculated.

Creating humor which will be later judged for funniness involves two steps: first, introducing a stimuli, and second, creating a funny response relevant to the stimuli presented. In the most common variation, researchers present participants with a picture or a cartoon with no caption and ask them to write a funny caption (Babad, 1974, Brodzinsky and Rubien, 1976, Eysenck, 1943, Feingold and Mazzella, 1991, Feingold and Mazzella, 1993, Kohler and Ruch, 1996, Koppel and Sechrest, 1970, Treadwell, 1970, Turner, 1980, Ziller et al., 1962). Other tasks seek a verbal response from a verbal stimulus. For example, McGhee (1974) presented children with absurd riddles such as "Why did the old man and his wife drive to the North Pole?" and asked them to provide funny answers that were later judged for funniness. Similarly, Shiloh (1982) asked participants to answer unusual questions in a funny way, such as "What would have happened if the oceans were full of orange juice?". Another example is to ask participants to provide a funny definition to a nonexistent term, such as *Yoga Bank* and *Fruit Jar*, or to write a funny ending to a social scenario that sets up funny responses (Christensen et al., 2018, Nusbaum et al., 2017). Only a few studies have sought visual output. For example, Howrigan and MacDonald (2008) asked participants to draw a picture of four animals (e.g., a giraffe), and four people with a specific profession (e.g., a professor).

Regardless of the type of stimuli and response employed, task administration methods are wide ranging. First, there is no standard time frame for how long each task requires to complete, ranging from as little as 30 s to write a funny caption for a captionless cartoon (e.g., Moran, Rain, Page-Gould, & Mar, 2014) to unlimited time (e.g., Feingold and Mazzella, 1993, Ziv, 1981b, Ziv, 1983), and anything in between. Second, the number of stimuli introduced to participants, and the number of responses allowed for each stimulus, vary substantially. For example, some researchers may ask for only one response per stimuli (e.g., Babad, 1974, Saroglou, 2002, Saroglou and Jaspard, 2001) while others will not restrict the number of captions the participant can produce (e.g., Greengross and Miller, 2011, Kohler and Ruch, 1996).

Evaluating individual differences in funniness also requires two steps. The first step is to ask judges to assess the level of funniness for each of the responses created by the participant. There is no consensus on how to evaluate and score individual humor production. With a few exceptions (e.g., Freiheit et al., 1998, Kozbelt and Nishioka, 2010), most judges of the humor stimuli are college students (both undergraduate and graduate) and professors (e.g., Nusbaum et al., 2017). The number of judges also differ, from 2 to as many as 81 (Amir and Biederman, 2016, Mickes et al., 2012). Most often judges are asked to rate the stimuli for "funniness", but in

some case judges are tasked with rating the "humorousness" (e.g., Saroglou, 2002, Saroglou and Jaspard, 2001,) or "wittiness" (e.g., Kohler & Ruch, 1996) of the caption or joke. How much each judge rates varies as well. In some cases, every judge evaluates the humor for all humor outputs, but sometimes, judges assess only a portion of the overall humor produced, mostly because there are too many stimuli for one person to reasonably handle. The rating scales themselves also vary, from merely a dichotomous distinction of whether the stimulus was funny or not (Saroglou and Jaspard, 2001, Saroglou, 2002, Ziv, 1983), to ratings ranging from 1 to 10 (Ruch, Beermann, & Proyer, 2009), and anything in between. In sum, there is much variability in the identity of judges, how many stimuli they rate from the overall sample, and what they are asked to do, when evaluating humor creativity.

The second part in evaluating humor is to create an overall score for HPA. As with other aspects of measuring humor ability, researchers use many methods to get an overall rating of funniness. For starters, some researchers use the raw scores of judges' ratings as the basis for their analyses (e.g., Moran et al., 2014), and others standardize the scores (e.g., Greengross & Miller, 2011). Furthermore, when more than one humor output is measured, and there are multiple judges rating the humor, there is a need to summarize all the scores to one statistic. Some researchers average all the responses across raters (e.g., Brodzinsky & Rubien, 1976), others take the highest score from each judge and average them (e.g., Greengross & Miller, 2011), and still others use many-facet Rasch models to distill the faceted design to a single score (e.g., Nusbaum et al., 2017).

In summary, there is no standardized procedure for assessing HPA. Nonetheless, while there is much variation in the tasks, the premises and procedures tend to follow the same mechanism. Overall, different measures of humor following the four-step process described in Fig. 1 are strongly correlated with each other (Christensen et al., 2018, Howrigan and MacDonald, 2008, Nusbaum et al., 2017).

1.3. The current research

It is clear that there is a pervasive stereotype that men have higher humor production abilities than women. Regardless of the reasons for such a belief, or why men and women might differ in their HPA, the veracity of sex differences in humor ability has not been systematically evaluated to date. Our goal is to create the first meta-analytic review on this topic. While a few qualitative reviews on sex differences in HPA exist, their conclusions were inconsistent with each other, and the scope of the reviews is limited (Greengross, 2014, Kotthoff, 2006, Martin, 2014). Thus, we had three main goals for our meta-analysis. First, we aimed to gather all available data on sex differences in HPA, and create the largest database on the topic to date. Second, we planned to estimate quantitatively the magnitudes of sex differences in HPA based on weighed effect sizes, from all available data. Third, we tested the possible influence of various moderators on sex differences in HPA (see below).

2. Method

2.1. Literature search

A key feature of the literature on humor ability is that relatively few studies have looked at and reported information about sex differences in HPA. Many publications that evaluated humor ability directly do not consider sex as an important factor of interest, and often do not report sex-specific data on HPA beyond the number of men and women participants. Other times, analyses of sex or gender differences are included, but the information is too limited to extract an effect size (e.g., Martin & Lefcourt, 1983). In either case, data on men's and women's humor ability might still exist and could be recovered, as sex is a common variable recorded in most studies. Thus, to avoid missing potentially relevant data, we searched for any study that had a measure of HPA, regardless of whether the information on sex differences was provided in the publication or not. If the researchers did not include data on men's and women's humor ability, or there was no information or analyses that could be reverse engineered to extract an effect size, we contacted the authors to attempt to retrieve the relevant data.

We employed multiple strategies to locate all relevant studies and data. First, we used a backward search (Card, 2015, p. 49-50) and examined the reference lists of key review publications that covered either sex or gender differences in humor or humor production in general. These reviews included Greengross, 2014, Kaufman et al., 2008, Lampert and Ervin-Tripp, 1998, Martin, 2014, Nusbaum, 2015, O'Quin and Derks, 1997; and Ruch (2008). Second, we used forward search (Card, 2015, p. 50-51) to look for studies that cited key articles on humor production, or ones that introduced new measures of HPA. These articles are: Babad, 1974, Brodzinsky and Rubien, 1976, Feingold and Mazzella, 1991, Kohler and Ruch, 1996, Koppel and Sechrest, 1970, Masten, 1986, Treadwell, 1970; and Turner (1980). Third, we emailed the listserv of the *International Society for Humor Research* (ISHS) asking for any unpublished data. Fourth, we searched the conference proceeding of the American Psychological Association, Association for Psychological Science, and ISHS, for any humor paper or poster presented since 2000, where data were available. Fifth, we searched the following databases: PsycINFO, PsycARTICLES, Web of Science, PubMed, ProQuest Dissertation & Theses Global, National Library of Israel, and Google Scholar. For searches in the databases we included the following phrases and their combinations: *humor ability*, *humor creat**, *humor product*, and cartoon caption**. All of the above searches were also performed using the non-American spelling, *humour*. Sixth, we contacted prominent researchers in the field of humor studies that come from non-English speaking countries, and asked if they knew about any publications in their field. In total, our literature searches included the following languages: Chinese, English, French, German, Hebrew, Hungarian, and Japanese. Publications in Hebrew were translated by the first author, and an author of a Hungarian paper translated all the relevant data we requested. Seventh, all authors of the current paper had unpublished data that was relevant to the meta-analysis, and knew about other researchers with additional unpublished data that we were able to obtain. We read the abstract, the full text, or both to determine if they contained relevant data for our meta-analysis. We concluded our search in August 2018.

2.2. Inclusion and exclusion criteria

Our approach was to include studies of humor production tasks that comprise of the two elements that we believe are the most essential for assessing true individual differences in humor ability with minimum biases: one, the creation of a new humor output (i.e., not a completion of a known joke or a joke recall), and two, the evaluation of the humor produced by independent

judges who do not know the identity or any characteristics of the humor producer. Thus, to be included in the meta-analysis, studies had to include the following criteria:

(1) The study must have included a sample of men and women.

(2) The humor production task must have included creation of verbal humor.

(3) Participants must have generated spontaneous new or innovative humor as part of the humor production task. Studies that were based on self-reports, or scales such as the Multidimensional Sense of Humor Scale (MSHS) (Thorson & Powell, 1993), or studies where participants had to complete a joke from multiple possible responses, given the setup, were excluded.

(4) Judges had to be blind to any characteristic of the humor producers. This excludes studies in which judges observed participants in the lab or on video (e.g., Inglis, Zach, & Kaniel, 2014).

(5) Judges rated the humor for funniness. Studies in which there were no judges, and the ratings of HPA were solely based on counting the number of responses or humor attempts, rather than actual ratings of the humor produced, were excluded (e.g., Hall, 2015). One study where judges were instructed to rate responses based on consensual sense of humor, i.e., if the answers could be seen as funny in principle by most people, regardless of the judges' own appraisal (essentially evaluating humor attempts rather than actual humor), was excluded (Hull, Tosun, & Vaid, 2017).

A couple of exclusion criteria were also applied. Studies on children (preadolescence) were excluded (e.g., Masten, 1986). In addition, studies that were conducted on participants with brain damage were not considered, though no such study met all the other inclusion criteria.

In cases where a study met all the inclusion criteria, but lacked sufficient information for a calculation of an effect size, we contacted the authors to obtain the information. A total of 22 authors were contacted, and eight of them provided data that could be included in our meta-analysis. In a few cases, the paper was too old and the data were lost, and one author refused to share the data. Additionally, the authors of the present paper had three unpublished datasets and knew of one more unpublished dataset that met our criteria. All these data were procured and included in the meta-analysis (see Fig. 2 for the PRISMA schematic screening of the studies) (Moher, Liberati, Tetzlaff, & Altman, 2009).

Any disagreements about what studies should be included were discussed and resolved among the authors. Altogether, 28 studies met our inclusion criteria, with 36 independent samples, and a total sample of 5057 participants (1677 men, 3380 women, 67% women).

2.3. Coding procedure

All studies were coded by two independent coders, one of whom was the first author. All disagreements were discussed and resolved between the coders. For each of our samples, we estimated the effect sizes based on standardized mean differences (Cohen's *d*) (Cohen, 1988). Most calculations were based on means, SD's and samples sizes, but when they were not available, we employed inferential statistics to calculate the effect size (see Card, 2015). These

procedures included the use of *F*-statistics, *t*-statistics, and *p*-values. We elected to use Cohen's *d* as our estimate for the standardized mean difference over Hedges' *g* as most samples were at least modest in size, and results based on Hedges' *g* were nearly identical to those based on Cohen's *d* (Card, 2015, p. 90-91). A positive sign of Cohen's *d* denotes a higher humor ability for men.

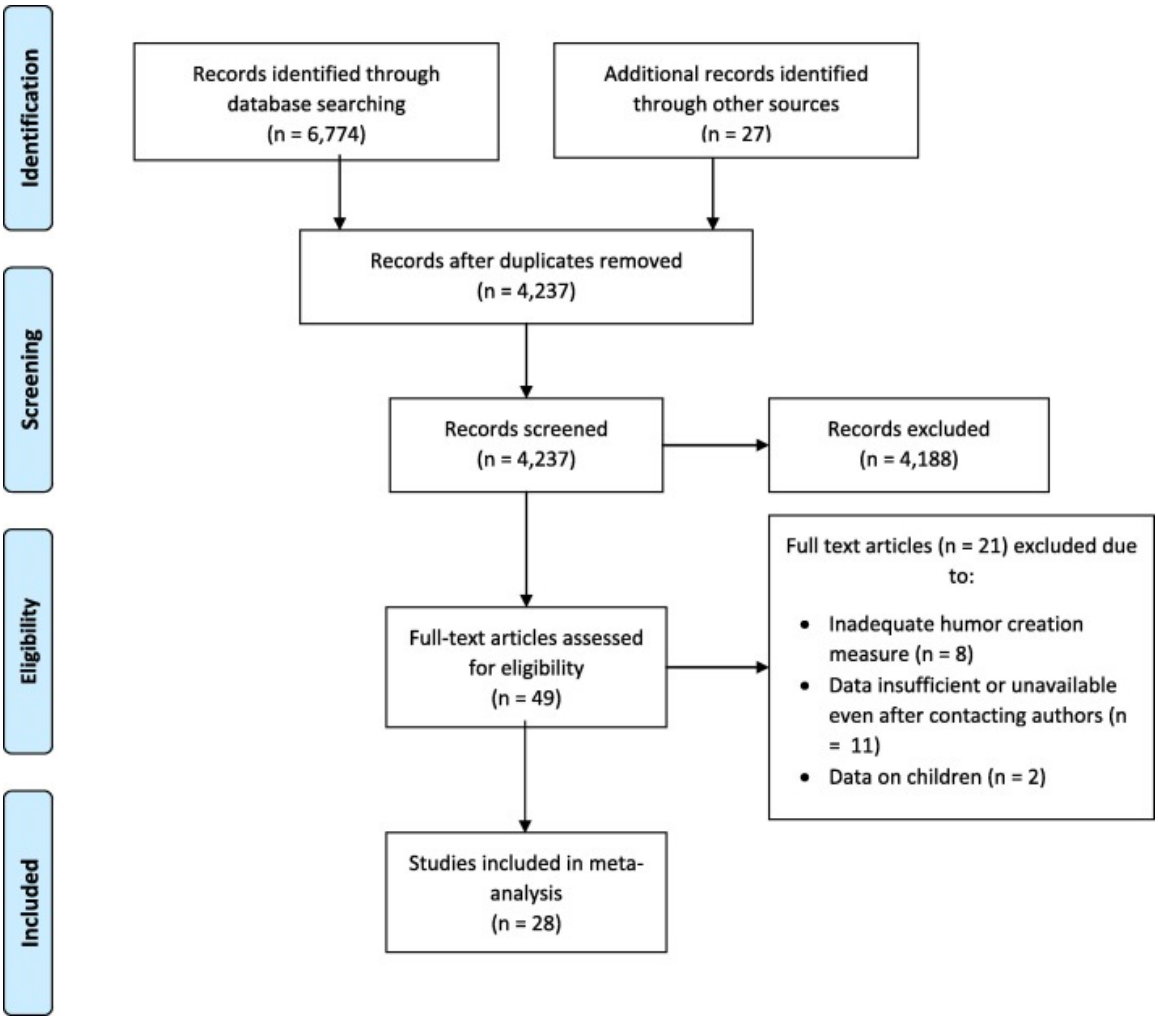

Figure 2. PRISMA flow chart of screening process of studies selected for the meta-analysis.

Several studies included more than one measure of HPA per sample. In such cases, we averaged the effect sizes to produce one *d* effect size for each sample. In total, we calculated average effect size for 12 out of the 36 samples with multiple outcomes. One exception was Mollica (1983), where we chose one effect size from the two reported. We believe this is justifiable as the effect size was based on the averages across all captions produced, and the one excluded was computed based on the average of the first caption only. Both measures are highly correlated with each other, and the overall average score embodies within it the average for the first caption.

Note that the procedures used to produce the overall HPA scores for each study vary, as there is no consensus on the best way to calculate a humor score. As discussed earlier, researchers

employ various protocols to calculate a humor score, which are partially dependent on the tasks employed, and whether one or more stimuli are used to generate humorous responses. For example, when multiple stimuli are used, researchers may average all responses across raters (Brodzinsky & Rubien, 1976), or only average the best outputs across raters (Greengross & Miller, 2011). For single tasks, there is no need to average responses, but they might be combined with other humor production tasks to create one overall score (e.g., Nusbaum et al., 2017). Still, despite the variability in the scoring of HPA, evidence suggests that different methods and scoring procedures yield similar results (Christensen et al., 2018, Mollica, 1983, Nusbaum et al., 2017).

2.4. Coding of potential moderators

We identified several variables that could potentially moderate sex differences in humor ability. The aim of the moderators was twofold. First, we wanted to test possible biases in the studies or the data collected. Some of these potential biases are associated with research on sex differences, such as the sex of the researcher, while other possible biases may relate to study characteristics, such as the type and nature of the samples or the country in which data were collected. Publication bias was also a concern, and we included several variables that specifically address this possibility (see below). Second, we included moderators that could illuminate or provide new insights into the source of sex differences in HPA, if it exists. These moderators pertain mostly to the measures, procedures, and evaluations associated with the humor ability task. It is possible that some moderators may serve both functions and both help detect biases and also elucidate the nature of sex differences in HPA. For example, *publication year* can both illustrate how sex differences in HPA has varied over the years, but also indicate changes in measurements of HPA across time.

We coded for 26 variables that were intended to be used as moderators, but not all were included in the moderator analyses, either due to little variability in the outcome, or because very few studies reported the relevant information. The list of the six excluded moderators and the justification for their exclusion appear in Table 1. In total, 20 variables served as moderators.

Table 1. Excluded moderators from the meta-analysis.

Moderator	Reason for exclusion
Department of first author	Lack of variability: 78% of samples were from a psychology department
Language of publication	Lack of variability: 92% of samples were in English
Judges' identity	Lack of variability: 89% of humor was rated by students and professors
Humor fluency	Limited data: only 28% of studies reported this variable
Reliability of judges' ratings	Lack of variability: 80% of the samples with Cronbach's α reported a reliability of 0.7 or higher
Sex differences in judges' ratings	Limited data: only 14% of studies reported this variable

Note. Humor fluency measures men's and women's number of responses for studies where participants were allowed to produce an unlimited number of humor outputs. For the reliability variable, various reliability measures were calculated, with 43% of the samples reporting Cronbach's α, and 31% of samples using other reliability measures.

Publication year. Publication year could help identify possible changes and trends of sex differences in HPA over time.

Affiliation of first author. The country in which the first author's affiliation appears on the manuscript, or in the case of unpublished data, where the researcher resides. This categorical variable includes the following countries/regions: North America (US & Canada), Europe (Austria, Belgium, Germany, Hungary, United Kingdom), and Israel. We grouped together the European countries as there were too few studies for each country to be included as a separate category. This variable may roughly represent the country in which the data were collected (i.e., where participants came from), but not in all cases (see the *sample country* variable). This moderator allows us to test whether the results are consistent cross-culturally.

Sex of first author. We dichotomously coded whether the first author was a man or a woman. This moderator aims to test for possible sex bias by the researcher conducting the study.

Single-sex team. This is a complementary, dichotomous variable to the *sex of first author* with yes/no coding, also intended to gauge a possible bias in publications with authors consisting of only one sex.

No. of authors. A larger number of authors may be less susceptible to biases.

Publication status. This is a dichotomous variable with peer-reviewed/not peer-reviewed coding. Note that for the peer-reviewed publications, the actual data on sex difference in HPA may not have been reported (e.g., was not paramount to the study), but there was still sufficient information to conclude that such data exist, and could be obtained. This moderator is one of the most common methods to assess publication bias.

Data availability in the peer-reviewed publication. This moderator aims to examine possible biases within peer-reviewed publications using yes/no coding that denotes whether sufficient data to calculate an effect size was included in a peer-reviewed publication (when applicable). In several cases, the peer-reviewed study did not contain the relevant data, though such data existed. The relevant information was obtained by contacting the authors directly.

Overall data publication status. This dichotomous moderator codes whether sufficient data on sex differences was included in a peer-reviewed publication, compared to all other manuscripts. Coding of 'yes' marks that data were published in a peer-reviewed paper and available for analysis. Coding of 'no' means that the data were either not included in a peer-reviewed publication but could still be accessed, or that it was not peer-reviewed (thesis, dissertation or unpublished data). This moderator could be viewed as a more precise estimate of publication bias, as it distinguishes between data that were fully reported in a peer-reviewed publication and all other data.

Sample group. This is a dichotomous variable with college students/non-college categories. We grouped the non-student samples under one category as they included many different groups (high school students, professional and amateur stand-up comedians and comedy writers, candidates for a tour guiding course abroad, online participants, and adolescent inpatients). Two studies that included a mix of students with other participants were excluded from the sample group moderator analysis.

Sample country. Similar to the *affiliation of first author* moderator, this categorical variable included the following countries/regions: North America (US & Canada), Europe (Austria, Belgium, Germany, Hungary, United Kingdom), Israel, and Worldwide. These comparisons will allow to test if there are country specific effects.

Mean sample age. This quantitative variable aimed at testing whether effects are consistent across different age means. For studies reporting a range of the participants' ages, the middle point was used to denote their mean age.

Sample size. The total number of participants is used to estimate publication bias. When publication bias is present, smaller sample sizes are associated with larger effect sizes.

Humor creation task. Most humor creation tasks employ the cartoon/picture caption paradigm, with no other task being common enough to comprise a category of its own. Therefore, 'new caption' denotes studies where participants were introduced with a captionless cartoon or picture, and were instructed to produce funny captions, and 'other' indicates studies with all other humor creativity tasks. The 'other' category included tasks based on verbal stimuli, in which participants were asked to complete a joke or a sentence in a funny way, write a funny story, write a funny resume or profile, write a funny definition for an absurd term, narrate a film in a funny way, or a composite score based on various non-caption stimuli. In a few cases, participants were asked to produce new captions in addition to other tasks and thus were coded as 'both'. This moderator can shed light on whether the type of the production humor task has any effect on sex differences in HPA.

No. of humor creation tasks. This is a quantitative variable denoting the number of humor creation items that served as a stimuli. More items may allow for more flexibility in the responses, and to express higher quality of humor.

No. of responses per task. This dichotomous variable coded 'one' vs. 'multiple' compares humor creation tasks that require the participants to produce only one funny outcome per each stimuli vs. studies which ask to produce multiple outcomes (usually unlimited number). The moderator allows us to compare humor tasks that limit participants to their best possible funny output, to those which give participants the freedom to try various attempts at being funny.

No. of levels in funniness scale. Most researchers use a numeric scale (e.g., 1–7), but for the studies using a dichotomous variable (e.g. funny vs. non-funny) the variable was coded as two. This moderator could illuminate whether effect sizes are more or less pronounced depending on the number of levels in the scale.

Time limited. This dichotomous variable with yes/no categories indicates whether participants had a limited time to produce the humorous output, or were given unlimited time to do so. Time restriction may reduce the quality of the humor, as there is less time to think, and more pressure to produce. On the other hand, giving participants unlimited time to write funny responses may induce fatigue or boredom. These time effects may differ by sex, so we will use it as a moderator.

Average task time. This moderator complements the *time limited* variable and is more accurate in evaluating the effect of time on participant's performance by measuring in minutes the average time participants were allocated to complete the humor production task (only for time-limited tasks). The moderator was calculated by dividing the total task time by the number of tasks participants had to complete. Sex differences in HPA may be more or less pronounced as a function of the amount of time allocated for each task.

No. of judges. A higher number of judges are likely to create a more precise evaluation of the humor produced.

Male to female judge ratio. This is a quantitative variable representing the ratio of male to female judges. Values above 1 mean that there were more male raters than females. This moderator can help explain if sex differences in HPA are due to disproportional number of male or female judges.

We used Comprehensive Meta-Analysis software to analyze the data (Borenstein, Hedges, & Rothstein, 2005), and our figures were produced with the JASP software (Team, 2018). We first describe our overall analytic approach, overview the studies and their effect sizes, and compute the combined effect. We then address the possibility of publication bias, and report results for all moderators' analyses.

We analyzed the data based on both fixed- and random-effects models and the results were comparable to each other, with similar conclusions. Thus, we only report the random-effects results, which allow a more generalized inference about the mean of a distribution of effect sizes, and not just a single effect size (Card, 2015, p. 230-256). Random-effects models also require fewer assumptions about the statistical model and yield more conservative estimates, and thus are usually preferred over fixed-models (Borenstein, Hedges, Higgins, & Rothstein, 2009, p. 61-86).

3. Results

Table 2, Table 3 present an overview of the studies included in the meta-analysis, along with the estimated effect sizes and moderators' codings. We estimated the overall effect size of sex difference in HPA across 36 independent samples ($N = 5057$). Results from the random-effects model analysis show a mean estimated $d = 0.321$, 95% CI [0.237, 0.405], $Z = 7.46$, $p < .0001$, see Fig. 3. Put in a different way, using the standard normal cumulative distribution function, the combined effect size of $d = 0.321$ indicates that 63% of men score above the mean HPA of women (Card, 2015, p. 124). Homogeneity test was significant, $Q(34) = 56.14$, $p < .02$, with $I^2 = 37.66$. This result indicates that effect sizes have a small to moderate amount of heterogeneity, i.e., not estimates of a single population value, thus adding further justification for the use of random-effects models.

Table 2. Overview of studies, effect sizes, and moderators included in the meta-analysis.

Study	Sample size: Men	Sample size: Women	Effect size (d)	SE	Rel. weight	Affiliation of First author[a]	Sex of first author	Single sex team	No. of authors	Publication status	Data availability in the PR publication	Overall data publication status
Amir and Biederman (2016)[b]												
Comedians	17	3	0.68	0.64	0.42	NAM	Male	Yes	2	PR	No	No
Controls	10	7	0.05	0.49	0.69	NAM	Male	Yes	2	PR	No	No
Brodzinsky and Rubien (1976)	40	44	0.48	0.22	2.59	NAM	Male	No	2	PR	Yes	Yes
Christensen et al. (2018)	38	232	0.08	0.18	3.52	NAM	Male	No	4	PR	No	No
Christensen and Silvia (2016)	42	139	0.66	0.18	3.41	NAM	Male	No	3	NPR	N/A	No
Edwards and Martin (2010)	92	123	0.23	0.14	4.53	NAM	Female	No	2	PR	Yes	Yes
Freiheit et al. (1998)												
Adolescent inpatients	23	32	1.13	0.29	1.69	NAM	Female	No	3	PR	Yes	Yes
High school students	43	42	0.19	0.22	2.66	NAM	Female	No	3	PR	Yes	Yes
Geher, Betancourt, and Jewell (2017)	26	65	0.27	0.23	2.41	NAM	Male	No	3	PR	No	No
Greengross and Miller (2011)	200	200	0.37	0.10	5.84	NAM	Male	Yes	2	PR	Yes	Yes
Greengross et al. (2017)	38	79	0.43	0.20	2.99	Europe	Male	No	3	NPR	N/A	No
Greengross et al. (2012)	42	8	0.15	0.39	1.07	NAM	Male	Yes	3	PR	No	No
Howrigan and MacDonald (2008)	70	150	0.40	0.15	4.29	NAM	Male	Yes	2	PR	Yes	Yes
Kaufman (2016)	103	642	0.41	0.11	5.62	NAM	Male			NPR	N/A	No
Kellner and Benedek (2017)	41	110	−0.09	0.18	3.33	Europe	Female	No	2	PR	No	No
Kim, Zeppenfeld, and Cohen (2013)	34	62	0.41	0.22	2.69	NAM	Female	No	3	PR	No	No
Kudrowitz (2010)	52	32	0.18	0.22		NAM	Male	Yes	1	NPR	N/A	No
Lehman, Burke, Martin, Sultan, and Czech (2001)	21	39	0.36	0.27	1.90	NAM	Female	No	5	PR	Yes	Yes
Mickes et al. (2012)	16	16	0.24	0.35	1.23	NAM	Female	No	5	PR	Yes	Yes
Mollica (1983)[c]	21	49	0.29	0.26	2.02	NAM	Male	Yes	1	NPR	N/A	No
Moran et al. (2014)	66	93	0.29	0.16	3.85	NAM	Male	No	4	PR	Yes	Yes
Nusbaum et al. (2017)												
Study 1	38	125	0.19	0.19	3.26	NAM	Female	No	3	PR	No	No
Study 2	44	116	0.33	0.18	3.44	NAM	Female	No	3	PR	No	No
Study 3	45	93	0.37	0.18	3.33	NAM	Female	No	3	PR	No	No
Renner and Manthey (2018)	83	254	0.18	0.13	4.91	Europe	Male	No	2	PR	No	No
Saroglou and Jaspard (2001)												
Humorous video	12	15	0.93	0.41	0.97	Europe	Male	No	2	PR	Yes	Yes
No video/control	8	21	0.35	0.42	0.93	Europe	Male	No	2	PR	Yes	Yes
Religious video	12	17	0.54	0.38	1.08	Europe	Male	No	2	PR	Yes	Yes

Study	Sample size: Men	Women	Effect size (d)	SE	Rel. weight	Affiliation of First author[a]	Sex of first author	Single sex team	No. of authors	Publication status	Data availability in the PR publication	Overall data publication status
Saroglou (2002)	18	54	0.89	0.28	1.80	Europe	Male	Yes	1	PR	No	No
Séra, Boda-Ujlaky, and Gyebnár (2015)	43	90	−0.07	0.19	3.28	Europe	Male	No	3	PR	Yes	Yes
Shiloh (1982)												
Study 1	36	57	−0.07	0.21	2.73	Israel	Female	Yes	1	NPR	N/A	No
Study 2	36	64	0.33	0.21	2.78	Israel	Female	Yes	1	NPR	N/A	No
Townsend (1982)	47	63	−0.13	0.19	3.12	NAM	Male	Yes	1	NPR	N/A	No
Ziv (1981b)[d]	162	182	0.69	0.11	5.45	Israel	Male	Yes	1	PR	Yes	Yes
Ziv (1983)												
Control[e]	28	30	0.17	0.26	2.01	Israel	Male	Yes	1	PR	Yes	Yes
Experiment	30	32	0.51	0.26	2.07	Israel	Male	Yes	1	PR	Yes	Yes

Note. Positive d denotes higher humor ability for men. Rel. weight = relative weight for the random model; NAM = North America; N/A = not applicable; PR = peer-reviewed; NPR = Non-peer-reviewed.
[a] For North American countries (NAM), all are from the US except Edwards and Martin (2010).
[b] Humor production was based on recall of captions that were generated under a fMRI scan. Following the scan, participants were asked to write down the same captions they thought about while in the machine.
[c] The dissertation included two effect sizes, but we calculated the effect size based on only the averages across all captions. The effect size based on the average of the first caption was excluded (see also coding procedures).
[d] The sd of 7.54 reported in Ziv (1981b) is probably wrong, as it is incongruent with women's sd of 17.28. Ziv (1981a) reports a sd of 17.54, which we used here.
[e] The study included non-significant results, therefore we used a conservative p-value of 0.51 to calculate the effect size.

Table 3. Moderators included in the meta-analysis.

Study	Sample group	Sample country[1]	Sample age mean	Humor creation task	No. of humor creation tasks	No. of responses per task	Funniness scale	Time limited	Average task time (minutes)	No. of judges	Male to female judge ratio
Amir and Biederman (2016)											
Comedians	NCS	NAM	32.05	Caption	Variable	Multiple	7	Yes	0.25	81	1.45
Controls[a]		NAM	24.90	Caption	Variable	Multiple	7	Yes	0.25	81	1.45
Brodzinsky and Rubien (1976)	CS	NAM		Caption	12	One	6	Yes	5.00	6	1.00
Christensen et al. (2018)	CS	NAM	19.08	Both	3	One	5	No	N/A	3	2.00
Christensen and Silvia (2016)	CS	NAM	19.10	Other	9		3	No	N/A	8	0.60
Edwards and Martin (2010)	CS	NAM	18.58	Both	5	Multiple	5	Yes	2.50	6	0.50
Freiheit et al. (1998)											
Adolescent inpatients	NCS	NAM	15.44	Caption	8	One	5	Yes	1.88	6	1.00
High school students	NCS	NAM	15.13	Caption	8	One	5	Yes	1.88	6	1.00
Geher et al. (2017)	NCS	World	26.32	Caption	2	One	5	No	N/A	3	0.50
Greengross and Miller (2011)	CS	NAM	20.60	Caption	3	Multiple	7	Yes	3.33	6	0.50[b]
Greengross et al. (2017)	CS[c]	Europe	21.94	Caption	3	Multiple	5	Yes	3.33	12	1.40

Study	Sample group	Sample country[1]	Sample age mean	Humor creation task	No. of humor creation tasks	No. of responses per task	Funniness scale	Time limited	Average task time (minutes)	No. of judges	Male to female judge ratio
Greengross et al. (2012)	NCS	NAM	35.46	Caption	3	Multiple	7	Yes	3.33	6	0.50
Howrigan and MacDonald (2008)	CS	NAM	22.00	Other	11	Multiple	7	No	N/A	4[d]	1.00
Kaufman (2016)	CS	NAM	24.19	Caption	4						
Kellner and Benedek (2017)	CS	Europe	23.10	Caption	6	Multiple	4	Yes	2.5	10	
Kim et al. (2013)	CS	NAM		Caption	5	Multiple		Yes	1.2	6	
Kudrowitz (2010)	NCS[e]	NAM	28	Caption	3	Multiple	3	Yes	1.67	12[d]	0.79
Lehman et al. (2001)	CS	NAM		Other	1		5			2	
Mickes et al. (2012)	CS	NAM		Caption	20	One	6	Yes	2.25	81	0.72
Mollica (1983)	CS	NAM		Caption	18	Multiple	6	Yes	2.22	5	0.67
Moran et al. (2014)	NCS	World	30.00	Caption	34	One	7	Yes	0.5	4	0.00
Nusbaum et al. (2017)											
Study 1	CS	NAM	19.00	Both	3	One	5	Yes	N/A	4	1.00
Study 2	CS	NAM	19.00	Both	3	One	5	Yes	2.00	5	0.67
Study 3	CS	NAM	18.70	Both	3	One	5	No	N/A	2	1.00
Renner and Manthey (2018)	CS[c]	Europe	33.17	Caption	6	Multiple	10	Yes	0.42	3	0.00
Saroglou and Jaspard (2001)											
Humorous video	CS	Europe		Caption	24	One	2			2	1.00
No video/control	CS	Europe		Caption	24	One	2			2	1.00
Religious video	CS	Europe		Caption	24	One	2			2	1.00
Saroglou (2002)	CS	Europe	23.42	Caption	24	One	2			2	1.00
Séra et al. (2015)[f]		Europe	29.50	Caption	6	Multiple	9	Yes	2.5	14	
Shiloh (1982)											
Study 1	NCS	Israel	17.50	Other	4	Multiple	2	Yes	1.00	30	
Study 2	NCS	Israel	34.53	Other	1	Multiple	2	Yes	4.00		
Townsend (1982)	NCS	NAM	17.00[g]	Caption	4	Multiple	6	Yes	0.75	2	
Ziv (1981b)	NCS	Israel	15.50[g]	Caption	10	One	6	No	N/A	3	
Ziv (1983)											
Control	NCS	Israel	15.50[g]	Caption	20	Multiple	2	No	N/A		
Experiment	CS	Israel	15.50[g]	Caption	20	Multiple	2	No	N/A		

Note. Blank indicates no available data; NCS = non-college students; CS = college students; NAM = North America; World = worldwide; Caption = new caption; Other = Verbal stimuli; Both = new caption and verbal stimuli; N/A = not applicable.

[a] Sample group moderator data were excluded as it included a mix of college and graduate students, as well as faculty members.

[b] The original paper mistakenly reported that there were four men and two women judges, where in fact there were four women and two men judges.

[c] Majority of participants were college students.

[d] Total number of judges was higher, but this was the number that judged each task.

[e] Majority of participants were non-college students.

[f] Sample group moderator data were excluded as half the participants were students and half not.

[g] Middle point was used for a range of ages.

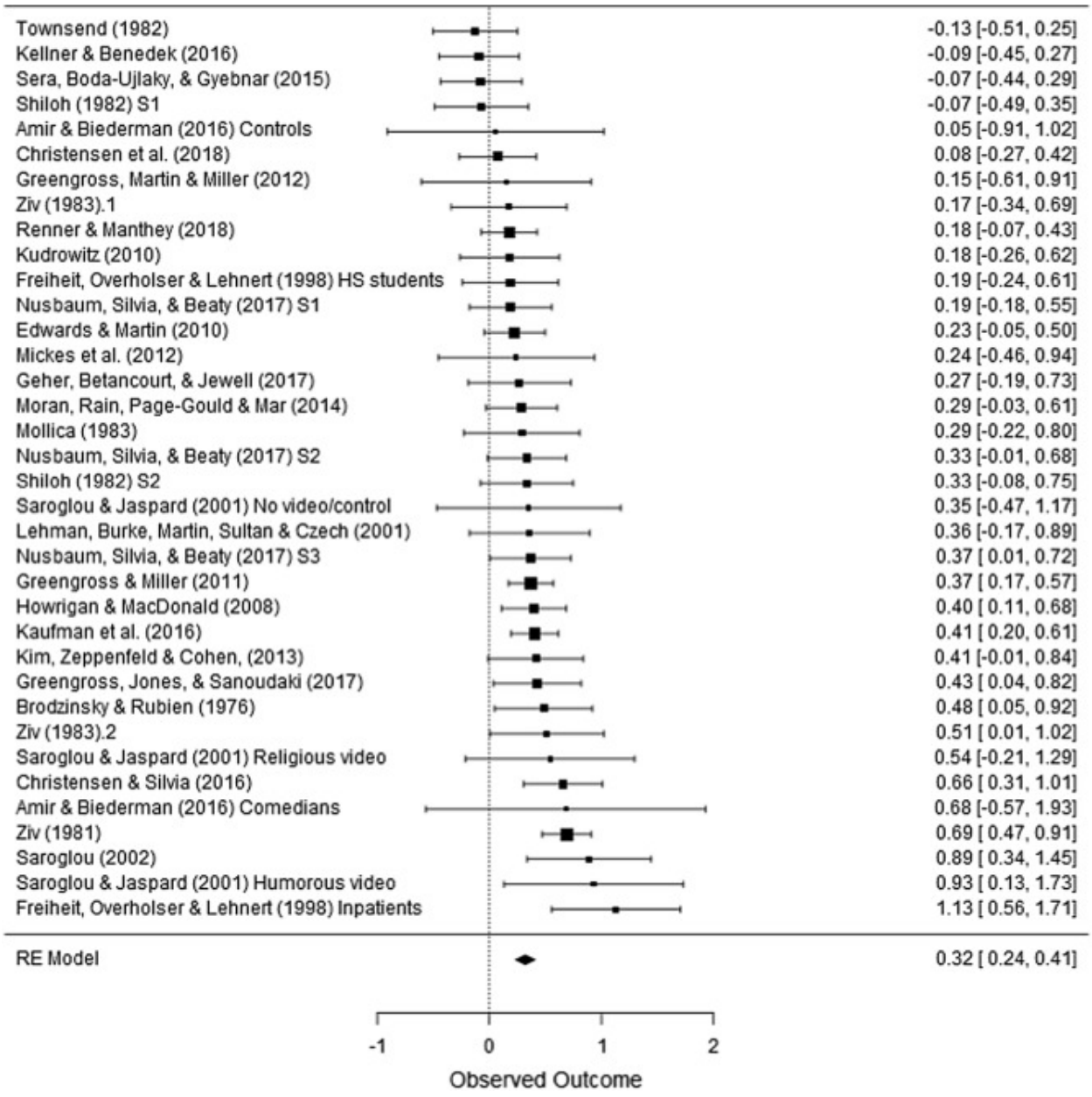

Figure 3. Forest plot displaying effect sizes with 95% confidence intervals and relative weights, as well as the combined effect. All estimates are based on a random-effect model.

Fig. 3 presents a forest plot of all independent effect sizes and the overall combined effect size, along with their respective 95% confidence intervals (CI), and each sample's relative weights. Effect size ranged from −0.13 to 1.13, with four negative effect sizes and 32 positive ones.

3.1. Assessing publication bias

We took several steps to reduce the risk of publication bias and to test for its existence. We first addressed the issue in the initial stages of the meta-analysis, during the literature search, in an attempt to minimize publication bias. Additionally, we report analyses that tested whether such bias exists after the data were collected.

3.1.1. Publication status

In an attempt to reach as many studies as possible and minimize publication bias, we did not explicitly search for sex or gender differences, and did not include the words ‘sex’ or ‘gender’ in our searches. It is possible that studies focusing on sex differences (and hence, include ‘sex’ or ‘gender’ as key words), are more likely to report that such differences exist. By using broader search terms, we were able to minimize this bias and still access many studies that contained relevant data on sex differences in HPA. These data were either ancillary to the main findings or were not reported at all. In total, 12 of the 29 peer-reviewed papers used in the meta-analysis did not include sufficient data on sex differences, but the data were later obtained from the authors. Overall, 20 of the 36 effect sizes in the current meta-analysis (61% of the participants) were not previously published, either coming from unpublished manuscripts (e.g. theses), or retrieved from authors of published papers that did not report the data. Additionally, from all 16 effect sizes published in peer-review publications, seven showed significant sex differences results and nine did not. Thus, whether sex differences in HPA existed did not seem to be a major factor determining the publication of the paper, minimizing the file drawer problem (Rosenthal, 1979).

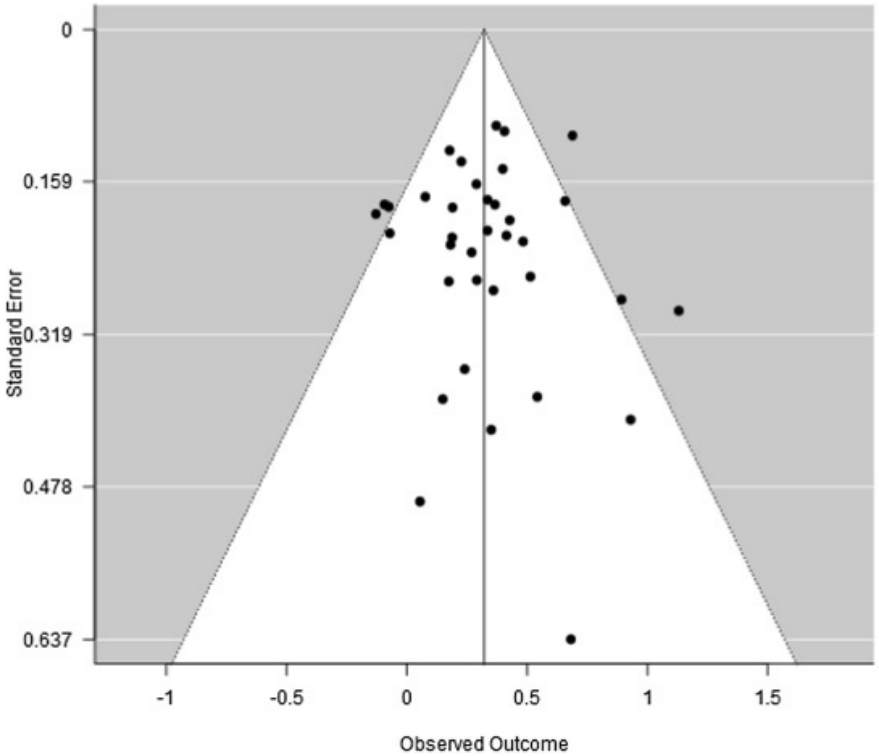

Figure 4. Funnel plot of effect sizes against their standard errors. The solid line indicates the combined effect size. White area represents 95% of CI**.**

3.1.2. Funnel plot

One of the most common ways to evaluate publication bias is by graphically displaying effect sizes against their standard errors through a funnel plot (Card, 2015, p. 263-266). Visual inspection of the funnel plot in Fig. 4 reveals a roughly symmetric distribution, with no obvious

outliers. The figure shows lower variability in effect sizes for larger samples, as one would expect if no publication bias exists. In addition to visually inspecting the funnel plot, we conducted two tests for the asymmetry of the funnel plot. The Begg and Mazumdar rank correlation test for funnel plot asymmetry tests the correlation between effect sizes and standard errors using Kendall's rank correlation coefficient (Begg & Mazumdar, 1994). Results show no evidence for asymmetry (*Tau* = 0.117, *Z* = 1.01, *p* = .31). The second test for the funnel plot asymmetry was Egger's regression test (Egger, Smith, Schneider, & Minder, 1997). As with the previous test, there was no indication for publication bias (*b* intercept = −0.028, 95% CI [-1.187, 1.129]), *t*(34) = 0.050, *p* = .96. Overall, these tests and the graphs reveal no evidence of publication bias due to small sample sizes having large effects.

3.1.3. Trim and fill analysis

The trim and fill method is a procedure to correct the estimate of the combined effect size by imputing potentially missing effect sizes to make the funnel plot symmetric (Duval & Tweedie, 2000). A new combined estimate is produced by adding the effect sizes that would fall left of the mean. This method identifies how many studies are missing and what their effect sizes are. However, our analysis revealed that no missing studies were identified and needed to be added, resulting in the same adjusted point estimate and confidence interval as the main results (*d* = 0.321, 95% CI [0.237, 0.405]). Thus, this analysis showed no evidence of publication bias.

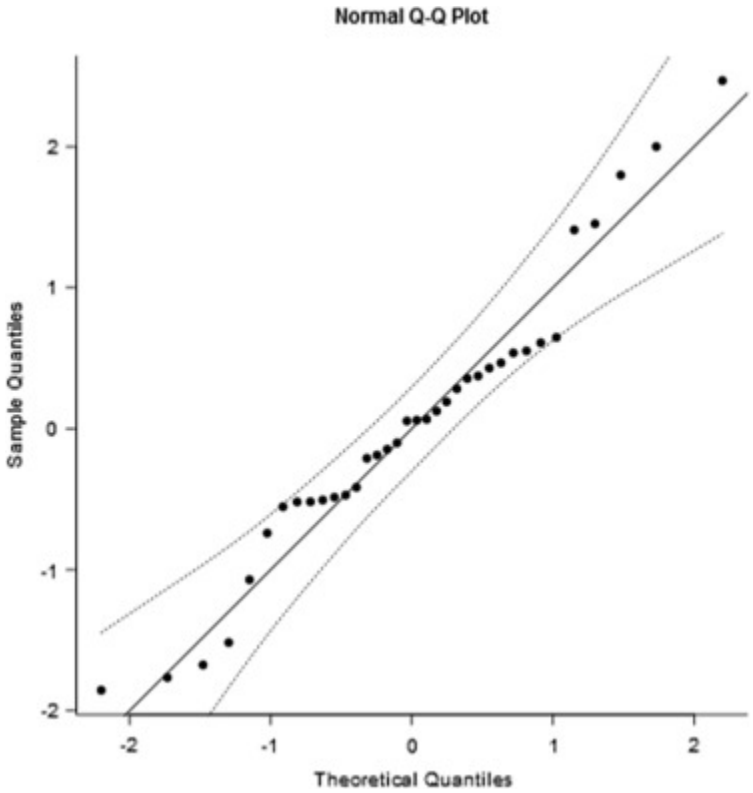

Figure 5. Normal quantile plot. The diagonal line represents normal distribution of the dots close to the diagonal line indicate normal distribution of effect sizes. Dashed lines mark 95% CI.

3.1.4. Normal Q-Q plot

Another graphic examination for publication bias is the normal quantile plot. In this plot, observed quantiles are plotted against the normal distribution quantiles. If the observed data are normally distributed, we would expect the points on the plot to fall roughly on a straight line (Wang & Bushman, 1998). Fig. 5 displays the Normal Q-Q plot. As can be seen from the graph, the points are roughly on a straight line, and all effects fall within 95% of the normal line. These results show no evidence of the data deviating from a normal distribution, or of publication bias.

3.2. Quantitative moderator analyses

Moderator analyses attempt to explain the heterogeneity of the effect sizes and identify variables contributing to a lower or higher combined effect size (Card, 2015, p. 198-228). To assess the impact of moderators on the combined effect size, we first ran a series of univariate meta-regressions for each of the nine quantitative moderator variables. In these tests, we predict the combined effect size from every moderator variable. All models included an intercept and test whether the coefficient is equal to zero. Table 4 reports the range of the results, the median, the coefficient b, 95% CI, *Z* scores, *Q* values, and *p*-values. As can be seen from the table, none of the moderators was significant (all tests are for *p*-value of 0.05). Most notably, the variable *sample size* is a commonly used moderator for testing publication bias. If publication bias exists, and the combined effect size is positive, we would expect to find a negative association between sample size and effect size, meaning that a small sample size produces large effect sizes (Card, 2015, p. 266). Results showed no association between sample size and effect size, thus there is no evidence for publication bias.

Table 4. Univariate meta-regressions for quantitative moderator analysis.

Moderator	K	Range	Median	b	95% CI	Z	Q(1)	p
Sample size	36	17–745	94.5	0.0001	[−0.0004, 0.0006]	0.42	0.18	0.672
Publication year	36	1976–2018	2010.5	−0.0026	[−0.0086, 0.0033]	−0.87	0.75	0.386
No. of authors	35	1–5	2	−0.0111	[−0.0966, 0.0743]	−0.26	0.07	0.799
Sample age mean	28	15.13–35.46	21.7	−0.0098	[−0.0264, 0.0069]	−1.15	1.32	0.250
No. of humor creation tasks	34	1–34	6	0.0089	[−0.0018, 0.0197]	1.63	2.64	0.104
No. of levels in funniness scale	34	2–10	5	−0.0261	[−0.0700, 0.0177]	−1.17	1.36	0.243
Average Task Time	21	15 s-5 min	2 min	0.0566	[−0.0298, 0.1429]	1.28	1.65	0.199
No. of judges	32	2–81	5.5	−0.0032	[−0.0101, 0.0037]	−0.91	0.82	0.365
Male to female judge ratio	25	0–2	1	0.0293	[−0.1435, 0.2021]	0.33	0.11	0.739

Though none of the moderator analyses was significant, we further visually inspected the distribution of every moderator against the effect sizes to see if any trends emerged. Three moderators had unusual range. The first, number of humor creation tasks, ranged from 1 to 34, with a quarter of studies administering more than 16 tasks. However, the moderator had no discernible correlation with effect size. Second, there were four studies with 30 or more judges, but their effect size ranged from −0.07 to 0.68, showing no particular trend. Third, funniness scales ranged from 2 to 10. As with the previous two moderators, there was no apparent association between scale's range and effect sizes.

3.3. Categorical moderation analyses

Using mixed-effects categorical moderator analysis (Borenstein et al., 2009), we calculated effect sizes and 95% CI for each level of all 11 categorical moderators. Results are displayed in Table 5, along with all between group heterogeneity tests. Of all 11 variables tested, only one moderator, *number of responses per task*, was significant with $p = .012$. As three samples included in this variable had no available data, a further analysis directly comparing the 15 studies that used one response per task to the 18 studies that included multiple responses per task, was performed. Results were still significant ($Q = 5.213$, $p = .022$), meaning that studies that included humor creation tasks which required only one response per stimuli showed a larger sex difference in HPA ($d = 0.425$), compared to studies that allowed for multiple responses per humor creation task ($d = 0.220$), with both effect sizes still significant ($p < .0001$). Additionally, the moderator *time limited*, had a *p*-value of 0.056. Since it included one study that was both time limited and not (with two different tasks), and six studies where data were unknown, we decided to compare No and Yes levels directly. Analysis revealed no differences between these two groups ($Q = 3.323$, $p = .068$).

Table 5. Mixed-effects categorical moderator analysis.

Moderator/Group	k	Men	Women	d	95% CI	Q	p
Main effect	36	1677	3380	0.321	[0.237, 0.405]		
Affiliation of first author						0.067	0.967
North America	23	1130	2375	0.326	[0.247, 0.405]		
Europe	8	255	640	0.302	[0.053, 0.551]		
Israel	5	292	365	0.354	[0.049, 0.659]		
Sex of first author						0.823	0.364
Female	12	469	879	0.266	[0.122, 0.410]		
Male	24	1208	2501	0.348	[0.245, 0.451]		
Single sex team						0.813	0.666
No	21	805	1807	0.298	[0.193, 0.404]		
Yes	14	769	931	0.331	[0.175, 0.488]		
Unknown	1	103	642	0.405	[0.196, 0.614]		
Publication status						0.253	0.615
Not peer-reviewed	8	375	1125	0.280	[0.097, 0.463]		
Peer-reviewed	28	1302	2255	0.333	[0.236, 0.430]		
Availability in PR publication						2.816	0.093
No	12	436	1129	0.238	[0.114, 0.362]		
Yes	16	866	1126	0.392	[0.261, 0.523]		
Overall data publication status						2.201	0.138
No	20	811	2251	0.265	[0.160, 0.370]		
Yes	16	866	1126	0.392	[0.261, 0.523]		
Sample group						4.917	0.086
College students	21	1016	2580	0.337	[0.256, 0.418]		
Non-college students	13	608	703	0.316	[0.125, 0.506]		
Mixed	2	53	97	−0.059	[−0.399, 0.281]		
Sample country						0.175	0.981
North America	21	1038	2217	0.329	[0.240, 0.417]		
Europe	8	255	640	0.302	[0.053, 0.551]		
Israel	5	292	365	0.354	[0.049, 0.659]		
Worldwide	2	92	158	0.282	[0.022, 0.543]		
Humor creation task						1.321	0.517
New caption	26	1215	2242	0.337	[0.225, 0.449]		
Other	5	205	449	0.353	[0.125, 0.581]		
Both	5	257	689	0.235	[0.087, 0.384]		
No. of responses per task						8.792	0.012

Moderator/Group	k	Men	Women	d	95% CI	Q	p
One	15	591	1147	0.425	[0.277, 0.573]		
Multiple	18	920	1413	0.221	[0.128, 0.314]		
Unknown	3	166	820	0.460	[0.289, 0.630]		
Time limited?						7.542	0.056
No	8	441	923	0.422	[0.253, 0.591]		
Yes	21	1024	1544	0.239	[0.137, 0.340]		
Both	1	38	125	0.190	[−0.175, 0.555]		
Unknown	6	174	788	0.475	[0.305, 0.646]		

Note. PR = peer-reviewed; Other = Verbal stimuli; Both = new caption and verbal stimuli.

Following Borenstein et al. (2009) we planned to perform a multiple regression analysis only on the moderators that were found significant in the moderator analysis. However, as only one moderator was significant (number of responses per task), there was no need for further analysis. However, as the only significant result was found in one moderator out of 20 moderator tests, it is possible that such a result is reached by chance because of the multiple tests (Type I error). Thus, we need to interpret the results of this moderator with caution.

4. Discussion

The present meta-analysis provides the first comprehensive quantitative evaluation of sex differences in humor production ability. We sought to collect all available studies assessing new humor outputs (mostly verbal) rated for funniness by independent judges blind to any characteristic of the producer. We believe that these types of humor production tasks and the method by which they are evaluated reflect the most objective measure of true humor abilities. Results reveal a small to moderate effect, with men scoring higher than women.

We took several steps to address and minimize the possibility of publication and other potential biases. First, we used random-effects models and weighted means to estimate the combined effect size, a more conservative approach for analyzing the data (Borenstein et al., 2009, Card, 2015). Second, our statistical analyses and tests, together with visual inspections of funnel and Normal Q-Q plots, showed no indication of publication bias. Overall, more than 60% of the data (20 out of 36 samples) were unpublished, either from unpublished datasets, theses and dissertations, or based on peer-reviewed publications where the data did not appear in the paper. The latter consists of 12 samples (44% of the published data), and included studies that did not focus on sex differences in HPA. Thus, it is unlikely that the presence or absence of sex differences in HPA played any role in the decision to publish the study. In fact, the combined effect size for peer-reviewed publications was almost identical to that of not peer-reviewed studies (0.333 and 0.280 respectively). Moreover, given the debate surrounding sex differences in HPA, it is unclear what type of bias might have existed. Results suggesting that men and women do not differ in their HPA are as important and informative as data indicating that sex differences in HPA do exist. Indeed, roughly half of the peer-reviewed publications revealed no significant sex differences, while the other half showed men having higher HPA. Thus, it is unlikely that the reason an unpublished manuscript was not published had anything to do with the presence or absence of sex differences in HPA. In sum, all indications are that there is little evidence for publication bias, and our view is that such bias is unlikely.

Third, to address the possibility of additional biases, we included in our meta-analysis a large number of moderators that could potentially influence the results. One aim was to test whether there is any bias stemming from the authors' or judges' sex. We found no evidence for such biases. The sex of the first author, and whether a single or mixed sex team conducted the study, had no influence on the results. In addition, despite variation in the male to female judge ratio across studies, the overall number of male and female judges was nearly identical, 171 and 169 respectively (based on all studies that reported the figure, 26 out of 36 samples). A moderator test revealed no effect for this variable on the combined effect; thus, it seems unlikely that judges' sex had any impact on the results.

Only one of the variables yielded a significant effect on the estimated combined effect size. This moderator, *number of responses per task* suggests that asking participants to produce one humorous output may result in higher sex differences in HPA for men, compared with tasks that allow for multiple humour outputs. This finding may indicate that men might have a further advantage when asked to produce their best humor once, while women may be funnier when given the chance to create multiple responses. Still, it is important to remember that in either scenario, men still scored significantly higher than women.

Though our moderator analyses did not reveal other statistically significant moderators' effects, some results may still illuminate important trends in the sources of sex differences in HPA. In particular, an important aspect of all humor creation tasks is whether the participants produced humor under time constraints or not. The results showed that having unlimited time to complete the task was associated with larger sex differences in HPA, compared to tasks restricted in time. Having unlimited time to create humor may have reduce the stress involved in creating humor, and allow participants more time to think about the task, something that benefited men more than women. However, it is also important to note that for all studies where participants were limited in the amount of time to produce humor, more time was not associated with larger sex differences. In other words, just knowing that the time is limited seemed to have some effect on the magnitude of sex differences.

It was also interesting to find some null moderator effects. For example, sex differences remained similar in more than 40 years of research, though there are relatively few studies prior to 2000 for which the data were available. Similarly, the results were surprisingly similar across cultures and samples, with college students and non-college students showing almost identical sex differences. Nonetheless, these results are provisional, as researchers may find different results in samples that are more diverse in the future.

4.1. Explaining sex differences in higher humor production ability

Sex differences in HPA found in this study may reflect evolved sex differences in mating preferences and strategies that were shaped by sexual selection theory (Darwin, 1871). Women's higher parental investment and larger reproduction costs than men make them choosier when selecting a mate, and more attentive to traits that could result in higher fitness (Buss, 2016, Trivers, 1972). HPA could be one such trait, as people vary in their ability to produce humor, and it is a reliable, hard to fake signal of intelligence, a highly desirable trait that increases fitness and serves as a mental fitness indicator to attract mates (Christensen et al.,

2018, Greengross and Miller, 2011, Howrigan and MacDonald, 2008, Kaufman et al., 2011, Miller, 2000a, Miller, 2000c). As women are choosier than men, we would expect women to be more sensitive and attuned to men's display of high HPA. As a consequence, a stronger intra-sexual competition among men ensues, resulting in an overall higher average HPA for men (Bressler and Balshine, 2006, Bressler et al., 2006, Miller, 2000a, Miller, 2000b). Based on this logic, when selecting a mate, men should use humor more often and more creatively to attract women and signal their mate value, while women should be more sensitive to men producing high quality humor. Various research supports this theory, and the view that HPA is valued differently and divulges disparate information for men and women. Compared to men, choosier women value humor as a more important trait when selecting a mate, while men make more effort to impress women and advertise their humor ability, including in real ecological settings, such as dyadic conversations and on dating sites (Lippa, 2007, Provine, 1993, Sprecher and Regan, 2002, Todosijević et al., 2003, Wilbur and Campbell, 2011). Women also prefer a man with higher HPA, while men are more attracted to a woman that laughs at their humor, rather than a woman with high HPA, as smiles and laughter signal the woman may have a romantic interest in them (Bressler and Balshine, 2006, Bressler et al., 2006, Hone et al., 2015). Viewing HPA as a mental fitness indicator relies on the connection between HPA and intelligence (Miller, 2000a, Miller, 2000c), and numerous studies have shown positive correlations between the two attributes (Christensen et al., 2018, Greengross and Miller, 2011, Howrigan and MacDonald, 2008, Kellner and Benedek, 2017).

Nevertheless, while a universal phenomenon, humor varies across cultures and reflects societal norms (T. Jiang, Li, & Hou, 2019). Little is known about sex differences in any facet of humor among non-Western populations, thus, the universality of sex differences in HPA found in our meta-analysis should be taken with caution. Most of our data come from WEIRD (Western, Educated, Industrialized, Rich, and Democratic) countries, which may delineate only a fraction of all human populations (Henrich, Heine, & Norenzayan, 2010b). Often, findings that are true for WEIRD samples do not replicate in non-WEIRD populations (Gurven et al., 2013, Henrich et al., 2010b, Henrich et al., 2010a). It is possible that non-WEIRD countries use and experience humor differently, which could influence the direction and magnitude of sex differences in HPA. If we want to draw conclusions that will be applicable to all humans, as the evolutionary explanation suggests, we need information about more diverse populations. For example, research suggests that East Asian people, such as from China and Taiwan, laugh less, and view themselves as less funny, compared to Western cultures such the United States and Canada (Jiang, Yue, & Lu, 2011). Western countries generally tend to value humor more, perceive humor as a more socially desirable trait, and view humorous people more positively, compared to Eastern countries (Yue, Jiang, Lu, & Hiranandani, 2016).

It is possible that the observed difference between men's and women's HPA is an artifact of the fact that certain types of humor, the ones that are considered of high quality, are more freely expressed by men than by women. For example, men may feel no restrictions in telling sexual and aggressive jokes, while women may be more inhibited in the use of these types of humor. However, some research suggests that while women are less likely than men to tell jokes in general, when they do, women are just as likely as men to use sexual and aggressive themes (Johnson, 1991). Mickes et al. (2012) compared the themes that men and women used when producing humor using the cartoon captioning task. Their results showed that when generating

the humorous captions, men produced significantly more sexual humor and used more profanity than women, though the overall usage of such humor was low (4.30% for men, 1.95% for women). More importantly, the use of sexual humor and profanity did not give men any advantage and did not contribute to their total higher HPA ratings compared to those of women. Also, neither men nor women judges rated these types of humor as funnier. The authors concluded that the higher humor ability of men could not be attributed to the use of sexual humor and profanity, but to other factors.

It is also important to note that despite the belief held by many that women enjoy sexual and aggressive humor less than men, reviews of the literature show mixed support to such claims (Lampert and Ervin-Tripp, 1998, Martin, 2014). Earlier research found women to enjoy sexual and aggressive humor less than men (Lampert and Ervin-Tripp, 1998, Martin, 2014), however, many of the jokes and cartoons used in these studies portray women as the target of the jokes, and the jokes tended to be sexist. When the targets of the jokes are men, or the jokes are not sexist or have neutral themes, women and men express similar levels of appreciation to the humor (Lampert & Ervin-Tripp, 1998). In our analysis there is little evidence to suggest that men and women judges evaluated the humor produced by either sex differently. Five studies in our meta-analysis tested for sex differences between judges' ratings, and only one found a significant difference (Mickes et al., 2012). Mickes et al. (2012) reported that both male and female raters judged men's HPA as higher than women's, but male raters gave male participants slightly higher ratings than female raters. In contrast, four other studies that tested for sex differences in judges' rating did not find any significant differences (Brodzinsky and Rubien, 1976, Greengross and Miller, 2011, Greengross et al., 2017, Greengross et al., 2012). Given the small number of studies to date, the role of judges clearly deserves more attention in future research.

There is some evidence that sexual and aggressive stimuli used to elicit the humor production may have an effect on the overall magnitude of sex differences in HPA. Brodzinsky and Rubien (1976) asked participants to produce spontaneous new humor in response to captionless cartoons that contained either sexual, aggressive, or neutral themes. Sex differences in HPA were found for sexual and aggressive themes, but not for the neutral cartoons. With the exception of Brodzinsky and Rubien (1976), no other study in our analysis reported whether the stimuli were sexual or aggressive in nature.

Another possible explanation for the observed difference lies in the nature of the task itself. The typical HPA task requires a crisp, focused response. Some research suggests that men are more likely to tell jokes, while women prefer telling funny stories and anecdotes (Crawford & Gressley, 1991). Although most tasks included in the meta-analysis do not exactly imitate a traditional canned joke structure, they are fairly constrained and do not afford longer, narrative-oriented responses. In addition, joke-telling is relatively a small fraction of everyday use of humor, where most humor arises during spontaneous social interactions (Martin and Kuiper, 1999, Provine, 2000). Varying the types of contexts of HPA tasks seems like a particularly promising direction for future work.

4.2. Limitations and suggestions for future research

As with any meta-analysis, there is always a possibility that more data exists. In fact, we know with certainty about several studies that matched our inclusion criteria, but for which the data could not be retrieved. Many researchers included measures of HPA in their study but did not report the relevant information, mostly because they did not focus on studying sex differences in humor. Most of these data come from older studies, and the raw data are now lost. Nonetheless, as our analysis on publication bias revealed, adding more data is unlikely to change the overall results. Still, we welcome any new studies on this topic, perhaps with new methodologies for measuring HPA.

Our meta-analysis included studies that used verbal humor as their main measure of HPA. It is possible that by focusing on non-verbal humor, the results would have been different. Another limitation is that the measures of HPA included in our meta-analysis are somewhat artificial, and do not represent everyday production of humor. Requesting people to produce humor on demand is challenging, and perhaps disadvantaged women more than men. It also ignores the social context in which most humor is produced (Provine, 2000), context that if taken into account may benefit women more than men. Perhaps sex differences in HPA vary depending on the environment in which it is produced. For example, women may have equal HPA scores as men when interacting with other women. Thus, studying various dyadic interactions of men and women in more ecologically valid situations, such as natural conversations, is crucial for fully understanding when and how sex differences in HPA emerge. Relatively few researchers conduct these types of studies (Hall, 2015, Provine, 1993, Provine, 2000). Still, humor is largely a social phenomenon and most humor is created in a social context while interacting with other people. Studying humorous interactions in the lab (Hall, 2015), or observing them in natural settings (Provine, 1993, Provine, 2000) should be a fruitful endeavor that requires more of our effort.

Another limitation are the ages of participants included in our meta-analysis. The samples contained somewhat restricted ages, ranging from 15 to 35, with a median age of 21.7 (see Table 4). Clearly, such samples are not representative of the whole population, but they do represent individuals at peak reproductive age. At these ages, following sexual maturity, people are at peak fertility. This is the period when the competition over mates is the strongest (Buss, 2016). As a result, and due to women's choosiness, it might represent the time when men try to impress women with their humor the most, thus resulting in higher HPA than women, as our meta-analysis found. Hence, results might be different for younger or older populations, with different theoretical implications.

4.3. Conclusion

The research presented here focused on one specific aspect of humor that is largely under-investigated in humor research, humor production ability. Despite finding men to have higher humor creation abilities than women on verbal humor, this difference should not be seen as representative of other types of humor, including non-verbal humor production ability. In fact, for most aspects of humor, men and women seem to exhibit many similarities, with relatively few differences (Martin, 2014). In regard to humor production abilities, the topic of sex differences is often reduced to blunt assertions such as that "Women are not funny" (e.g., Hitchens, 2007). We hope that our meta-analysis will help advance a more nuanced

discussion on the topic based on a systematic evaluation of the available scientific data. Examination of such data suggest that regardless of the underlying source of variability, men exhibit higher humor ability than women on the kinds of verbal tasks included in our sample of studies. It is important to remember that though robust, these differences are small to medium in size, and are based on averages. They do not reflect individual abilities, as both men and women vary largely in their abilities to produce humor. We tried to illuminate possible sources for the differences in HPA, what they might mean, theoretical implications, considerations for future research, and limitations. Humor is an important experience for most people, one that is largely unique to humans. We hope that our results will further foster the study of humor, advance theories pertaining to understanding and explaining sex differences in humor and other cognitive abilities, as well as foster research on humor ability.

5. Footnotes

The study was not preregistered. Data will be shared in a public repository before publication.

The authors received no financial support for the research, authorship, and/or publication of this article.

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Acknowledgments

We thank the following researchers for providing unpublished data and other information included in our meta-analysis: Ori Amir, Mathias Benedek, Judit Boda-Ujlaky, Dov Cohen, Glenn Geher, Varda Inglis, Olivia Jewell, Scott Barry Kaufman, Barry Kudrowitz, Karl-Heinz Renner, Vassilis Saroglou, and Shoshana Shiloh.

Author contribution

Conception of the study – GG, PS, EN. Searching for studies, data collection, coding: GG, PS. Data analysis and interpretation: GG, PS, EN. Drafting the article: GG. Critical revisions and final approval of the article: GG, PS, EN.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jrp.2019.103886. This is open data under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

References

References marked with an asterisk indicate studies included in the meta-analysis

Alexander, R. D. (1986). Ostracism and indirect reciprocity: The reproductive significance of humor. Ethology and Sociobiology, 7(3–4), 253–270. https://doi.org/10.1016/0162-3095(86)90052-X.

*Amir, O., & Biederman, I. (2016). The neural correlates of humor creativity. Frontiers in Human Neuroscience, 10(597). https://doi.org/10.3389/fnhum.2016.00597.

Antonovici, L., & Turliuc, M.-N. (2017). Humour and mate selection in a Romanian sample. Journal of Experiential Psychotherapy, 20(1), 38–45.

Archer, J. (1996). Sex differences in social behavior: Are the social role and evolutionary explanations compatible?. American Psychologist, 51(9), 909–917. https://doi.org/10.1037/0003-066X.51.9.909.

Babad, E. Y. (1974). A multi-method approach to the assessment of humor: A critical look at the humor tests. Journal of Personality, 42, 618–631. https://doi.org/10.1111/j.1467-6494.1974.tb00697.x.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. Biometrics, 1088–1101. https://doi.org/10.2307/2533446.

Borenstein, M., Hedges, L. J. H., & Rothstein, H. (2005). Comprehensive Meta-Analysis Version 3. Biostat, Englewood, NJ.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). Introduction to meta-analysis. Chichester, UK: John Wiley & Sons.

Bressler, E., & Balshine, S. (2006). The influence of humor on desirability. Evolution and Human Behavior, 27(1), 29–39. https://doi.org/10.1016/j.evolhumbehav.2005.06.002.

Bressler, E., Martin, R. A., & Balshine, S. (2006). Production and appreciation of humor as sexually selected traits. Evolution and Human Behavior, 27(2), 121–130. https://doi.org/10.1016/j.evolhumbehav.2005.09.001.

*Brodzinsky, D. M., & Rubien, J. (1976). Humor production as a function of sex of subject, creativity, and cartoon content. Journal of Consulting and Clinical Psychology, 44(4), 597–600. https://doi.org/10.1037/0022-006X.44.4.597.

Buss, D. M. (1988). The evolution of human intrasexual competition: Tactics of mate attraction. Journal of Personality and Social Psychology, 54(4), 616–628. https://doi.org/10.1037/0022-3514.54.4.616.

Buss, D. M. (2016). The Evolution of Desire: Strategies of Human Mating (4 edition). New York, NY: Basic Books.

Card, N. A. (2015). Applied meta-analysis for social science research. New York, NY: Guilford Publications.

Chafe, W. (1987). Humor as a disabling mechanism. American Behavioral Scientist, 30(1), 16–26. https://doi.org/10.1177/000276487030003003.

*Christensen, A. P., & Silvia, P. J. (2016). Unpublished data.

*Christensen, A. P., Silvia, P. J., Nusbaum, E. C., & Beaty, R. E. (2018). Clever people: Intelligence and humor production ability. Psychology of Aesthetics, Creativity, and the Arts, 12(2), 136–143. https://doi.org/10.1037/aca0000109.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (second edition). Hillsdale, NJ: Routledge.

Crawford, M., & Gressley, D. (1991). Creativity, caring, and context: Women's and men's accounts of humor preferences and practices. Psychology of Women Quarterly, 15, 217–231. https://doi.org/10.1111/j.1471-6402.1991.tb00793.x.

Darwin, C. (1871). The Descent of Man, and Selection in Relation to Sex. London: John Murray.

Davila-Ross, M., Owren, M. J., & Zimmermann, E. (2009). Reconstructing the evolution of laughter in great apes and humans. Current Biology, 19(13), 1106–1111. https://doi.org/10.1016/j.cub.2009.05.028.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. Biometrics, 56(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x.

Eagly, A. H., & Wood, W. (1999). The origins of sex differences in human behavior: Evolved dispositions versus social roles. American Psychologist, 54(6), 408–423. https://doi.org/10.1037/0003-066X.54.6.408.

Eaton, A. A., & Rose, S. (2011). Has dating become more egalitarian? A 35 year review using Sex Roles. Sex Roles, 64(11–12), 843–862. https://doi.org/10.1007/s11199-011-9957-9.

*Edwards, K. R., & Martin, R. A. (2010). Humor creation ability and mental health: Are funny people more psychologically healthy? Europe's Journal of Psychology, 6 (3), 196–212. https://doi.org/10.5964/ejop.v6i3.213.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. BMJ, 315(7109), 629–634. https://doi.org/10.1136/bmj.316.7129.469.

Eysenck, H. J. (1943). An experimental analysis of five tests of ‘‘appreciation of humor.”. Educational and Psychological Measurement, 3(1), 191–214. https://doi.org/10.1177/001316444300300119.

Feingold, A. (1992). Gender differences in mate selection preferences: A test of the parental investment model. Psychological Bulletin, 112(1), 125–139. https://doi.org/10.1037/0033-2909.112.1.125.

Feingold, A., & Mazzella, R. (1991). Psychometric Intelligence and Verbal Humor Ability. Personality & Individual Differences, 12(5), 427–435. https://doi.org/10.1016/0191-8869(91)90060-O.

Feingold, A., & Mazzella, R. (1993). Preliminary validation of a multidimensional model of wittiness. Journal of Personality, 61(3), 439–456. https://doi.org/10.1111/j.1467-6494.1993.tb00288.x.

*Freiheit, S. R., Overholser, J. C., & Lehnert, K. L. (1998). The association between humor and depression in adolescent psychiatric inpatients and high school students. Journal of Adolescent Research, 13(1), 32–48. https://doi.org/10.1177/0743554898131003.

Gallup, G. G., Ampel, B. C., Wedberg, N., & Pogosjan, A. (2014). Do orgasms give women feedback about mate choice?. Evolutionary Psychology, 12(5). https://doi.org/10.1177/147470491401200507.

Gamble, J. (2001). Humor in Apes. HUMOR: International Journal of Humor Research, 14(2), 163–179. https://doi.org/10.1515/humr.14.2.163.

*Geher, G., Betancourt, K., & Jewell, O. (2017). The link between emotional intelligence and creativity. Imagination, Cognition and Personality, 37(1), 5–22. https://doi.org/10.1177/0276236617710029.

Goddard, C., & Wierzbicka, A. (2004). Cultural scripts: What are they and what are they good for?. Intercultural Pragmatics, 1(2), 153–166. https://doi.org/10.1515/iprg.2004.1.2.153.

Goodwin, R. (1990). Sex differences among partner preferences: Are the sexes really very similar?. Sex Roles, 23(9), 501–513. https://doi.org/10.1007/BF00289765.

Greengross, G. (2014). Male production of humor produced by sexually selected psychological adaptations. In V. A. Weekes-Shackelford & T. K. Shackelford (Eds.), Evolutionary perspectives on human sexual psychology and behavior (pp. 173–196). New York, NY: Springer.

Greengross, G., Jones, M., & Sanoudaki, E. (2017). Enjoyment and production of humour among bilingual speakers. Unpublished data.

*Greengross, G., Martin, R. A., & Miller, G. F. (2012). Personality traits, intelligence, humor styles, and humor production ability of professional stand-up comedians compared to college students. Psychology of Aesthetics, Creativity and the Arts, 6, 74–82. https://doi.org/10.1037/a0025774.

*Greengross, G., & Miller, G. F. (2011). Humor ability reveals intelligence, predicts mating success, and is higher in males. Intelligence, 39, 188–192. https://doi.org/10.1016/j.intell.2011.03.006.

Gurven, M., Von Rueden, C., Massenkoff, M., Kaplan, H., & Lero Vie, M. (2013). How universal is the Big Five? Testing the five-factor model of personality variation among forager–farmers in the Bolivian Amazon. Journal of Personality and Social Psychology, 104(2), 354–370. https://doi.org/10.1037/a0030841.

Hall, J. A. (2015). Sexual selection and humor in courtship: A case for warmth and extroversion. Evolutionary Psychology, 13(3), 1–10. https://doi.org/10.1177/1474704915598918.

Halpern, D. F. (2011). Sex differences in cognitive abilities (4th ed.). New York, NY: Psychology press.

Halpern, D. F., Benbow, C. P., Geary, D. C., Gur, R. C., Hyde, J. S., & Gernsbacher, M. A. (2007). The science of sex differences in science and mathematics. Psychological Science in the Public Interest, 8(1), 1–51. https://doi.org/10.1111/j.1529-1006.2007.00032.x.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010a). Most people are not WEIRD. Nature, 466, 29. https://doi.org/10.1038/466029a.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). The weirdest people in the world? Behavioral and Brain Sciences, 33(2/3), 61–135. https://doi.org/10.1017/S0140525X0999152X.

Hitchens, C. (2007). Why women aren't funny. Vanity Fair, 557, 54–59.

Hone, L. S. E., Hurwitz, W., & Lieberman, D. (2015). Sex differences in preferences for humor: A replication, modification, and extension. Evolutionary Psychology, 13(1), 167–181. https://doi.org/10.1177/147470491501300110.

Hooper, J., Sharpe, D., & Roberts, S. G. B. (2016). Are Men Funnier Than Women, or do we just think they are?. Translational Issues in Psychological Science, 2(1), 54–62. https://doi.org/10.1037/tps0000064.

*Howrigan, D. P., & MacDonald, K. B. (2008). Humor as a mental fitness indicator. Evolutionary Psychology, 6(4), 652–666. https://doi.org/10.1177/147470490800600411.

Hull, R., Tosun, S., & Vaid, J. (2017). What's so funny? Modelling incongruity in humour production. Cognition and Emotion, 31(3), 484–499. https://doi.org/10.1080/02699931.2015.1129314.

Hurley, M., Dennett, D., & Adams, R. Jr., (2011). Inside jokes: Using humor to reverse-engineer the mind (1 ed.). Cambridge, MA: The MIT Press.

Hyde, J. S. (2005). The gender similarities hypothesis. American Psychologist, 60(6), 581–592. https://doi.org/10.1037/0003-066X.60.6.581.

Hyde, J. S. (2014). Gender similarities and differences. Annual Review of Psychology, 65, 373–398. https://doi.org/10.1146/annurev-psych-010213-115057.

Inglis, V., Zach, S., & Kaniel, S. (2014). Humor creator and the audience-a multidimensional model supported by in-vivo methodology. American Journal of Educational Research, 2(7), 503–512. https://doi.org/10.12691/education-2-7-12.

JASP Team (2018). JASP (Version 0.8.6) [Computer software].

Jiang, F., Yue, X. D., & Lu, S. (2011). Different attitudes toward humor between Chinese and American students: Evidence from the Implicit Association Test. Psychological Reports, 109(1), 99–107. https://doi.org/10.2466/09.17.21.PR0.109.4.99-107.

Jiang, T., Li, H., & Hou, Y. (2019). Cultural differences in humor perception, usage, and implications. Frontiers in Psychology, 10(123). https://doi.org/10.3389/fpsyg.2019.00123.

Johnson, A. M. (1991). Sex differences in the jokes college students tell. Psychological Reports, 68(3), 851–854. https://doi.org/10.2466/pr0.1991.68.3.851.

*Kaufman, S. B. (2016). Unpublished data.

Kaufman, S. B., DeYoung, C. G., Reis, D. L., & Gray, J. R. (2011). General intelligence predicts reasoning ability even for evolutionarily familiar content. Intelligence, 39(5), 311–322. https://doi.org/10.1016/j.intell.2011.05.002.

Kaufman, S. B., Kozbelt, A., Bromley, M. L., & Miller, G. (2008). The role of creativity and humor in human mate selection. In G. Geher & G. Miller (Eds.), Mating intelligence: Sex, relationships, and the mind's reproductive system (pp. 227–262). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

*Kellner, R., & Benedek, M. (2017). The role of creative potential and intelligence for humor production. Psychology of Aesthetics, Creativity, and the Arts, 11(1), 52–58. https://doi.org/10.1037/aca0000065.

*Kim, E., Zeppenfeld, V., & Cohen, D. (2013). Sublimation, culture, and creativity. Journal of Personality and Social Psychology, 105(4), 639–666. https://doi.org/10.1037/a0033487.

Kohler, G., & Ruch, W. (1996). Sources of variance in current sense of humor inventories: How much substance, how much method variance? HUMOR: International Journal of Humor Research, 9(3–4), 363–397. https://doi.org/10.1515/humr.1996.9.3-4.363.

Koppel, M. A., & Sechrest, L. (1970). A multitrait-multimethod matrix analysis of sense of humor. Educational and Psychological Measurement, 30(1), 77–85. https://doi.org/10.1177/001316447003000107.

Kotthoff, H. (2006). Gender and humor: The state of the art. Journal of Pragmatics, 38(1), 4–25. https://doi.org/10.1016/j.pragma.2005.06.003.

Kozbelt, A., & Nishioka, K. (2010). Humor comprehension, humor production, and insight: An exploratory study. HUMOR: International Journal of Humor Research, 23(3), 375–401. https://doi.org/10.1515/humr.2010.017.

*Kudrowitz, B. M. (2010). Haha and aha!: Creativity, idea generation, improvisational humor, and product design. (Doctoral dissertation). Massachusetts Institute of Technology, Boston.

Lampert, M. D., & Ervin-Tripp, S. M. (1998). Exploring paradigms: The study of gender and sense of humor near the end of the 20th century. In W. Ruch (Ed.), The sense of humor: Explorations of a personality characteristic (pp. 231–270). Berlin: Walter de Gruyter.

*Lehman, K. M., Burke, K. L., Martin, R., Sultan, J., & Czech, D. R. (2001). A reformulation of the moderating effects of productive humor. HUMOR: International Journal of Humor Research, 14(2), 131–161. https://doi.org/10.1515/humr.14.2.131.

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. Psychological Bulletin, 136(6), 1123–1135. https://doi.org/10.1037/a0021276.

Lippa, R. A. (2007). The preferred traits of mates in a cross-national study of heterosexual and homosexual men and women: An examination of biological and cultural influences. Archives of Sexual Behavior, 36(2), 193–208. https://doi.org/10.1007/s10508-006-9151-2.

Lippa, R. A., Collaer, M. L., & Peters, M. (2010). Sex differences in mental rotation and line angle judgments are positively associated with gender equality and economic development across 53 nations. Archives of Sexual Behavior, 39(4), 990–997. https://doi.org/10.1007/s10508-008-9460-8.

Martin, R. A. (2014). Humor and gender: An overview of psychological research. In D. Chiaro & R. Baccolini (Eds.), Gender and Humor: Interdisciplinary and International Perspectives. Berlin: Mouton de Gruyter.

Martin, R. A., & Kuiper, N. (1999). Daily occurrence of laughter: Relationships with age, gender, and Type A personality. HUMOR: International Journal of Humor Research, 12(4), 355–384. https://doi.org/10.1515/humr.1999.12.4.355.

Martin, R. A., & Lefcourt, H. (1983). Sense of humor as a moderator of the relation between stressors and moods. Journal of Personality and social Psychology, 45(6), 1313–1324. https://doi.org/10.1037/0022-3514.45.6.1313.

Masten, A. S. (1986). Humor and competence in school-aged children. Child Development, 57(2), 461–473. https://doi.org/10.2307/1130601.

McGee, E., & Shevlin, M. (2009). Effect of humor on interpersonal attraction and mate selection. Journal of Psychology: Interdisciplinary and Applied, 143(1), 67–77. https://doi.org/10.3200/JRLP.143.1.67-77.

McGhee, P. E. (1974). Development of children's ability to create the joking relationship. Child Development, 45(2), 552–556. https://doi.org/10.2307/1127988.

*Mickes, L., Walker, D., Parris, J., Mankoff, R., & Christenfeld, N. (2012). Who's funny: Gender stereotypes, humor production, and memory bias. Psychonomic Bulletin & Review, 19(1), 108–112. https://doi.org/10.3758/s13423-011-0161-2.

Miller, G. F. (2000a). The mating mind: How sexual selection shaped the evolution of human nature. New York, NY: Anchor books.

Miller, G. F. (2000b). Mental traits as fitness indicators: Expanding evolutionary psychology's adaptationism. In D. Lecroy & P. Moller (Eds.), Evolutionary perspectives on human reproductive behavior (pp. 62–74). New York, NY US: New York Academy of Sciences.

Miller, G. F. (2000c). Sexual selection for indicators of intelligence. In G. Bock, J. Goode, & K. Webb (Eds.), The nature of intelligence (pp. 260–275): Novartis Foundation Symposium 233. John Wiley.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. Annals of Internal Medicine, 151(4), 264–269. https://doi.org/10.1371/journal.pmed.1000097.

*Mollica, M. A. (1983). Paradox Recognition: A Proposed Common Cognitive Process Between Creativity And Humor (Doctoral dissertation). Ann Arbor: DePaul.

*Moran, J. M., Rain, M., Page-Gould, E., & Mar, R. A. (2014). Do I amuse you? Asymmetric predictors for humor appreciation and humor production. Journal of Research in Personality, 49, 8–13. https://doi.org/10.1016/j.jrp.2013.12.002.

Nevo, O., Nevo, B., & Yin, J. L. (2001). Singaporean humor: A cross-cultural, cross-gender comparison. Journal of General Psychology, 128(2), 143–156. https://doi.org/10.1080/00221300109598904.

Nusbaum, E. C. (2015). A meta-analysis of individual differences in humor production and personality. (Ph.D.). The University of North Carolina at Greensboro, Ann Arbor. ProQuest Dissertations & Theses Global database.

*Nusbaum, E. C., Silvia, P. J., & Beaty, R. E. (2017). Ha ha? Assessing individual differences in humor production ability. Psychology of Aesthetics, Creativity, and the Arts, 11(2), 231–241. https://doi.org/10.1037/aca0000086.

O'Quin, K., & Derks, P. (1997). Humor and creativity: A review of the empirical literature. In M. Runco (Ed.). Creativity research handbook (Vol. 1, pp. 223–252). Cresskill, NJ: Hampton Press.

Preuschoft, S., & Van-Hooff, J. A. (1997). The social function of 'smile' and 'laughter': Variations across primate species and societies. In U. Segerstrale & P. Molnar (Eds.), Nonverbal communication: Where nature meets culture (pp. 171–190). Mahwah, NJ: Lawrence Erlbaum Associates.

Provine, R. (1993). Laughter punctuates speech: Linguistic, social and gender contexts of laughter. Ethology, 95(4), 291–298. https://doi.org/10.1111/j.1439-0310.1993.tb00478.x.

Provine, R. (2000). Laughter: A scientific investigation. New York: Viking. Ramachandran, V. S. (1998). The neurology and evolution of humor, laughter, and smiling: The false alarm theory. Medical Hypotheses, 51(4), 351–354. https://doi.org/10.1177/147470490600400129.

*Renner, K.-H., & Manthey, L. (2018). Relations of dispositions toward ridicule and histrionic self-presentation with quantitative and qualitative humor creation abilities. Frontiers in Psychology, 9(78). https://doi.org/10.3389/fpsyg.2018.00078.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. Psychological Bulletin, 86(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638.

Ross, E., & Hall, J. A. (in press). The Traditional Sexual Script and Humor in Courtship. HUMOR: International Journal of Humor Research.

Ruch, W. (2008). Psychology of humor. In V. Raskin (Ed.). The primer of humor research (Vol. 8, pp. 17–100). Berlin, Germany: Mouton de Gruyter.

Ruch, W., Beermann, U., & Proyer, R. T. (2009). Investigating the humor of gelotophobes: Does feeling ridiculous equal being humorless? Humor-International Journal of Humor Research, 22(1–2), 111–143. https://doi.org/10.1515/HUMR.2009.006.

*Saroglou, V. (2002). Religiousness, religious fundamentalism, and quest as predictors of humor creation. International Journal for the Psychology of Religion, 12(3), 177–188. https://doi.org/10.1207/S15327582IJPR1203_04.

*Saroglou, V., & Jaspard, J.-M. (2001). Does religion affect humour creation? An experimental study. Mental Health, Religion & Culture, 4(1), 33–46. https://doi.org/10.1080/13674670010016756.

*Séra, L., Boda-Ujlaky, J., & Gyebnár, V. (2015). A Humorstílus és a kreativitás különbözo0 aspektusainak összefüggései [Relationships between different aspects of humor style and creativity]. Magyar Pszichologiai Szemle, 70(2), 295–312. https://doi.org/10.1556/0016.2015.70.2.1.

*Shiloh, S. (1982). Humor Creation: Its Relationship to Individual and Situational Characteristics. (Doctoral dissertation). Tel Aviv University, Tel Aviv, Israel.

Shlesinger, I. (2017). Girl Logic: The Genius and the Absurdity. New York, NY: Hachette Books.

Smith, J. E., Waldorf, V. A., & Trembath, D. L. (1990). Single white male looking for thin, very attractive.... Sex Roles, 23(11/12), 675–685. https://doi.org/10.1007/BF00289255.

Spelke, E. S. (2005). Sex Differences in Intrinsic Aptitude for Mathematics and Science?: A Critical Review. American Psychologist, 60(9), 950–958. https://doi.org/10.1037/0003-066X.60.9.950.

Sprecher, S., & Regan, P. C. (2002). Liking some things (in some people) more than others: Partner preferences in romantic relationships and friendships. Journal of Social and Personal Relationships, 19(4), 463–481. https://doi.org/10.1177/0265407502019004048.

Thorson, J. A., & Powell, F. C. (1993). Development and validation of a multidimensional sense of humor scale. Journal of Clinical Psychology, 49(1), 13–23. https://doi.org/10.1002/1097-4679(199301)49:1<13::AIDJCLP2270490103>3.0.CO;2-S.

Todosijevic´ , B., Ljubinkovic´ , S., & Aranc˘ic´ , A. (2003). Mate selection criteria: A trait desirability assessment study of sex differences in Serbia. Evolutionary Psuchology, 1, 116–126. https://doi.org/10.1177/147470490300100108.

Toro-Morn, M., & Sprecher, S. (2003). A cross-cultural comparison of mate preferences among university students: The United States vs. the People's Republic of China (PRC). Journal of Comparative Family Studies, 34(2), 151–170.

Townsend, J. W. J. (1982). Relationships among humor, creative thinking abilities, race, sex, and socioeconomic factors of advantagedness and disadvantagedness of a selected sample of high school students. (Doctoral dissertation). University of Georgia, Ann Arbor. ProQuest Dissertations & Theses Global database.

Treadwell, Y. (1970). Humor and creativity. Psychological reports, 26, 55–58. https://doi.org/10.2466/pr0.1970.26.1.55.

Trivers, R. L.(1972). Parental investment and sexual selection. In B. Campbell (Ed.), Sexual Selection and the Descent of Man: 1871–1971 (pp. 136–179). Chicago, IL: Aldine.

Turner, R. G. (1980). Self-monitoring and humor production. Journal of Personality, 48(2), 163–172. https://doi.org/10.1111/j.1467-6494.1980.tb00825.x.

Viana, A. (2017). Humour and laughter as vestiges of evolution. The European Journal of Humour Research, 5(1), 1–18. https://doi.org/10.7592/EJHR2017.5.1.viana.

Voyer, D., Postma, A., Brake, B., & Imperato-McGinley, J. (2007). Gender differences in object location memory: A meta-analysis. Psychonomic bulletin & review, 14(1), 23–38. https://doi.org/10.3758/BF03194024.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. Psychological Bulletin, 117(2), 250–270. https://doi.org/10.1037/0033-2909.117.2.250.

Voyer, D., Voyer, S. D., & Saint-Aubin, J. (2017). Sex differences in visual-spatial working memory: A meta-analysis. Psychonomic Bulletin & Review, 24(2), 307–334. https://doi.org/10.3758/s13423-016-1085-7.

Wang, M. C., & Bushman, B. J. (1998). Using the normal quantile plot to explore meta-analytic data sets. Psychological Methods, 3(1), 46–54. https://doi.org/10.1037/1082-989X.3.1.46.

Weisfeld, G. E. (1993). The Adaptive Value of Humor and Laughter. Ethnology and Social Biology, 14(2), 141–169. https://doi.org/10.1016/0162-3095(93)90012-7.

Wilbur, C. J., & Campbell, L. (2011). Humor in romantic contexts: Do men participate and women evaluate?. Personality and Social Psychology Bulletin, 37(7), 918–929. https://doi.org/10.1177/0146167211405343.

Yue, X., Jiang, F., Lu, S., & Hiranandani, N. (2016). To Be or Not To Be Humorous? Cross Cultural Perspectives on Humor. Frontiers in Psychology, 7(1495). https://doi.org/10.3389/fpsyg.2016.01495.

Zell, E., Krizan, Z., & Teeter, S. R. (2015). Evaluating gender similarities and differences using metasynthesis. American psychologist, 70(1), 10–20. https://doi.org/10.1037/a0038208.

Ziller, R. C., Behringer, R. D., & Goodchilds, J. D. (1962). Group creativity under conditions of success or failure and variations in group stability. Journal of Applied Psychology, 46(1), 43–49. https://doi.org/10.1037/h0045647.

*Ziv, A. (1981a). The psychology of humor. Israel: Yachdav.

*Ziv, A. (1981b). The self concept of adolescent humorists. Journal of Adolescence, 4, 187–197. https://doi.org/10.1016/S0140-1971(81)80038-3.

Ziv, A. (1983). The influence of humorous atmosphere on divergent thinking. Contemporary Educational Psychology, 8(1), 68–75. https://doi.org/10.1016/0361-476X(83)90035-8.