Cultural Sexism Moderates Efficacy of Psychotherapy:

Results from a Spatial Meta-Analysis

Maggi A. Price, PhD
School of Social Work, Boston College
Department of Psychology, Harvard University

John R. Weisz, PhD
Department of Psychology, Harvard University

Sarah McKetta, MSc
Department of Epidemiology, Columbia University Mailman School of Health

Jessie V. Ford, PhD
Department of Sociomedical Sciences, Columbia University Mailman School of Health

Micah R. Lattanner, PhD
Department of Epidemiology, Columbia University Mailman School of Health

Hilary Skov, BA
Department of Psychology, Harvard University

Elizabeth Wolock, BA
Department of Psychology, Harvard University

Mark L. Hatzenbuehler, PhD
Department of Psychology, Harvard University

Draft version 1 (9-17-2020). This paper has not yet been peer reviewed. Please do not copy or cite without the authors’ permission.
Abstract

We examined whether cultural sexism (county- and state-level gender attitudes) moderates the efficacy of psychotherapies by re-analyzing data from a previous meta-analysis of youth psychotherapy randomized controlled trials (4,233 effect sizes (ESs) from 319 studies; \(N=20,513\); ages 4-18). Higher cultural sexism was associated with lower ESs for studies with \(\geq 50\%\) girls; this association became stronger as the proportion of girls in the samples increased. Cultural sexism was unrelated to ESs for studies with >50% boys. An interaction between state- and county-level sexism revealed that psychotherapies were most beneficial when they were conducted in states and counties with the lowest levels of cultural sexism. Thus, the context in which psychotherapies are delivered is associated with psychotherapy efficacy for girls.

**Keywords:** cultural sexism, psychotherapy, treatment effectiveness, gender, children and adolescents

**Public Significance Statement:** This spatial meta-analysis found that psychotherapy randomized controlled trials with samples comprised of a majority of girls were significantly less effective in states with higher vs. lower levels of cultural sexism. These findings suggest that examining the social contexts in which psychotherapy interventions are delivered may yield new insights into who benefits most (and least) from mental health treatments.

**Trial registration:** PROSPERO identifier CRD42017072759
Cultural Sexism Moderates Efficacy of Psychotherapy:
Results from a Spatial Meta-Analysis

One of the most pressing issues in psychotherapy research is evaluating treatment effect heterogeneity (Bloom & Michalopoulos, 2013; Reardon & Stuart, 2017)—that is, identifying for whom, and under what conditions, mental health is improved by the interventions received. To date, studies have focused almost exclusively on the identification of individual (e.g., gender, race) and study-specific (e.g., treatment modality) characteristics that moderate intervention efficacy. Consequently, it remains largely unknown whether contextual features of the broader social environment in which therapies occur are associated with intervention impact (Johnson et al., 2017). The lack of research on contextual mechanisms that may affect therapy response is striking, given the robust literature that social/contextual factors—including income inequality (Pickett et al., 2006), social capital (Flores et al., 2017), neighborhood violence (Fowler et al., 2009), and community-level prejudice (Hatzenbuehler, 2016), among many others—shape mental health outcomes. The current paper seeks to address this gap in the literature by examining whether one potential contextual risk factor for mental health—cultural sexism—is associated with reduced intervention efficacy among girls receiving psychotherapy treatment. To address this research question, we draw on and integrate several relevant literatures from psychology, sociology, and public health—including gender as a social structure (Risman, 2004; 2017), feminist psychology (Brown, 2018), and macro-level stigma (Hatzenbuehler, 2016)—and utilize a recent innovation in studying contextual moderators of intervention efficacy: spatial meta-analysis (Johnson et al., 2017).

Psychopathology and Psychotherapy among Girls
Girls and women evidence higher rates of internalizing problems (e.g., depression, anxiety) relative to boys and men. Meta-analyses suggest that throughout adolescence and adulthood, girls are much more likely than boys to experience depression (Salk et al., 2017; Twenge & Nolen-Hoeksema, 2002), with adolescent girls and women being twice as likely to be diagnosed with major depression (Merikangas et al., 2011). Similar gender differences with respect to magnitude and chronicity have been found in studies of anxiety, with these disparities emerging even earlier (i.e., around age 6; Kessler et al., 2005). Although boys are significantly more likely than girls to be diagnosed with externalizing disorders (e.g., conduct disorder, attention deficit hyperactivity disorder (ADHD; Merikangas et al., 2011), this disparity is narrowing (Moffitt, 2003).

Studies of nationally-representative samples of youth indicate that boys are significantly more likely to use therapy compared to girls (Simpson et al., 2008), even in samples limited to youths with depression symptoms, externalizing problems (Zimmerman, 2005), and serious emotional and behavior difficulties (Simon et al., 2015). Several reasons have been posited, including that boys’ problems are more overt and thus more likely to be identified by adults who refer them to treatment (Gaub & Carlson, 1997). Despite differences in treatment utilization, outcomes do not appear to systemically differ across girls and boys in studies of internalizing treatment (Nilsen et al., 2013), though exceptions exist (e.g., Watson & Nathan, 2008).

There is a large literature on the causes of the gender disparity in internalizing psychopathology, with explanations emphasizing the multilevel interplay of biological, psychological, social, and macro-level (e.g., structural and cultural) factors (for a review, see Hatzenbuehler & McLaughlin, 2017). While empirical research has largely focused on identifying individual and interpersonal factors, feminist psychotherapy has long asserted that
contextual factors related to gender and gender inequality—including sex roles, gender socialization, and the stigmatization of girls and women—contribute to psychological difficulties among girls and women (Miller, 1976). In locating the causes of mental health problems in the larger social context, feminist psychotherapists therefore emphasize treatment components such as consciousness-raising, gender-role analysis, and social activism (Israeli & Santor, 2000).

**Gender as a Social Structure**

The theoretical arguments made in feminist psychotherapy are echoed in the sociological literature on gender as a social structure. In her seminal theory, Risman (2004) articulates the importance of an integrative, multi-level approach to the study of gender and sexism, with an explicit acknowledgment that individual, interactional, and macro levels are equally integral and complementary features of a larger social stratification system. Accordingly, gender inequality is produced at each level: through the development of gendered selves (the individual level); during interactions where boys/men and girls/women face distinct expectations, even when occupying similar status positions (the interactional level); and in institutional domains where resources and opportunities are distributed unevenly based on gender (the macro level).

In a recent revision to this theory, Risman (2017) stressed the importance of differentiating between cultural and material dimensions at each level of the gender structure. Cultural dimensions are conceptualized as hegemonic and ideological beliefs (macro level), interpersonal expectations and stereotypes (interactional level), and internalized gendered selves (individual level). Material dimensions, in contrast, include the distribution of resources (macro level), access to social networks (interactional level), and the physical body (individual level).

**Macro-Level Material Sexism**
Research on the physical and mental health consequences of macro-level sexism has focused largely on the material (vs. cultural) dimensions of the gender structure. These studies have leveraged variation in macro-level sexism across geographic units of analysis (e.g., countries, states, counties) using a variety of indicators of material sexism—which frequently include indices of women’s economic, political, and reproductive autonomy, both as absolute (e.g., the percentage of women living below the poverty line) and relative measures (e.g., the ratio of women to men living below the poverty line). This work has shown that women living in U.S. states with relatively higher macro-level material sexism—i.e., with relatively higher rates of women’s poverty, higher wage gaps, fewer legal protections, and less female representation in state government—are more likely to experience intimate partner violence, poor self-rated health, more chronic health conditions, poor physical functioning, and premature mortality (for a review, see King et al., 2020). County-level variation in macro-level material sexism in the U.S. has similarly been linked to differences in risk for female homicide (Vieraitis et al., 2016), child and infant homicide (Hunnicutt, 2007), and intimate partner violence (Gillespie & Reckdenwald, 2017).

Regarding mental health specifically, two studies have found that women residing in states with higher levels of macro-level material sexism—measured using four composite indices representing employment and earnings, political participation (e.g., women’s voter turnout and women’s representation in government), economic autonomy (e.g., women’s small business ownership), and reproductive rights—had higher rates of any mood disorder, depressive symptoms, major depression, and post-traumatic stress disorder (Chen et al., 2005; McLaughlin et al., 2011). While relatively less attention has been paid to the relationship between macro-level material sexism and girls’ health, one study found that higher state-level sexism (measured
using the same four composite indices as above) was related to higher rates of teen pregnancy (Koenen et al., 2006).

**Macro-Level Stigma and Mental Health**

This work has provided important advancements in understanding how the material dimensions that underlie macro-level sexism increase risk for poor outcomes among women, including psychopathology. Yet, there is a dearth of research focusing specifically on the mental health consequences of *cultural* dimensions of macro-level sexism—operationalized as ideological beliefs and gender norms—despite calls for this type of work (Risman, 2017; Salk et al., 2017). There is, however, an emerging body of empirical evidence that macro-level stigma, as measured via community-level ideologies and norms, can affect the mental health of low-status and marginalized groups (for reviews, see Hatzenbuehler, 2016; 2017a, b). In these studies, individuals’ attitudes (both implicit and explicit) related to particular stigmatized groups are aggregated to the community level—defined at various geographic scales (e.g., counties, states)—such that the level of stigma can be compared across communities. Aggregated social norms have been used in previous research to measure macro-level stigma across a range of groups, including Black people (e.g., Leitner et al., 2016), immigrants (e.g., Morey et al., 2017), lesbian, gay, and bisexual (LGB) populations (e.g., Hatzenbuehler et al., 2017), individuals with serious mental health problems (e.g., Evans-Lacko et al., 2012), and individuals living with HIV/AIDS (e.g., Miller et al., 2011; 2016).

This work has shown that macro-level stigma is associated with numerous adverse psychosocial and health outcomes among the stigmatized, including self-stigma (Evans-Lacko et al., 2012), disclosure concerns (Miller et al., 2011), symptoms of psychological distress (Miller et al., 2016), hypervigilance and hopelessness (Russell & Richards, 2003), perceived stress
(Rostosky et al., 2009), poor self-rated health (Hatzenbuehler et al., 2017b), and even shortened lifespans (Leitner et al., 2016). For instance, in one study, researchers measured macro-level stigma using aggregated data on public attitudes about people with mental health problems from 14 European countries. The study found that among people diagnosed with serious mental health problems, those living in countries with lower levels of stigma related to mental health problems reported lower rates of self-stigma and perceived discrimination than those in countries with higher levels of stigma (Evans-Lacko et al., 2012). In another study, researchers inferred macro-level stigma by collecting data on the share of votes for (or against) same-sex marriage in a national postal plebiscite on proposed same-sex marriage legislation in Australia. LGB people reported worse life satisfaction, lower social support, and poorer mental health outcomes in constituencies where the residents had higher (vs. lower) shares of voters against same-sex marriage, controlling for individual- and contextual-level factors (Perales & Todd, 2018).

**Spatio-Temporal Meta-Analyses**

Thus, emerging evidence indicates that macro-level stigma, measured via aggregated norms and ideologies, can negatively affect the mental health of low-status and marginalized groups. In the current study, we examine another potential way in which macro-level stigma and prejudice, as it specifically relates to sex/gender (referred to as cultural sexism, herein), may adversely affect mental health—namely, through undermining the efficacy of psychotherapy for girls. Addressing this research question is methodologically challenging, however, because most therapy studies are conducted in only one, or a few, communities. Limited variation in the social contexts of most studies precludes sensitive tests of whether contextual factors, such as cultural sexism, moderate intervention efficacy, because individuals are ubiquitously exposed to the same
environment, limiting the variation necessary to detect associations between contextual factors and intervention efficacy (should one exist).

Fortunately, recent advances in meta-analysis include a method for addressing this gap in the literature. Spatio-temporal meta-analyses allow for the measurement and statistical modeling of community contexts (e.g., prejudicial norms, economic inequality) and temporal dimensions (e.g., cultural trends) in relation to intervention efficacy (Johnson et al., 2017). Using this method to investigate whether particular social contexts explain heterogeneity of intervention effects may have important theoretical and practical implications. Spatio-temporal meta-analysis can be used to test conceptual models of the effects of social contexts on health, such as ecological (Bronfenbrenner, 1979) and social stress (Aneshensel & Phelan, 1999) theories. Practically, results from spatio-temporal meta-analyses may shed light on where interventions have (and have not) been conducted, and can suggest the development of new, adjunctive treatment modalities to enhance intervention efficacy in contexts where they are underperforming (Johnson et al., 2017).

In one of the first applications of spatio-temporal meta-analysis to behavioral and psychological interventions, Reid and colleagues (2014) re-analyzed a previously published meta-analytic database with information on effect sizes from 78 HIV prevention interventions aimed at improving condom use among African Americans. These individual-level interventions were geographically heterogeneous—that is, they took place across the U.S. in communities that differed in levels of macro-level racism, which was operationalized via measures of anti-Black attitudes and racial residential segregation. The interventions improved condom use among African Americans only in communities with low levels of macro-level racism (i.e., those with relatively positive attitudes toward African Americans and low levels of residential segregation).
Conversely, the interventions were ineffective when conducted in communities with the highest levels of racism (Reid et al., 2014). In addition, moderation analyses indicated that the association between macro-level racism and intervention efficacy was strongest among adolescents, a developmental period when social identities are still developing and when external evaluations, especially of marginalized statuses, may be especially salient (e.g., Gibbons et al., 2007).

**The Current Study**

The current study builds on the important contributions of this previous spatio-temporal meta-analysis by: 1) examining a different feature of the social context (i.e., cultural sexism); and 2) exploring a different intervention target (i.e., psychotherapies for youth). In addition, while prior work has tended to measure macro forms of stigma and sexism at a single level of analysis (e.g., countries: Pachankis et al., 2015; states: Homan, 2019; counties: Reid et al., 2014), the current study examined cultural sexism simultaneously across two geographic scales: states and counties. This dual measurement approach is important, because measuring cultural sexism at a single level of analysis (e.g., states only) may obscure heterogeneity across levels (i.e., more proximal environments can have attitudinal contexts that differ from the state level). Moreover, examining cultural sexism at more than one level permits the examination of cross-level interactions to determine whether cultural sexism at the state and county level interact to predict intervention efficacy. Because we examined cultural sexism at more than one level, it was necessary to measure it in similar ways across states and counties such that their association with intervention efficacy could be compared. Consequently, we operationalized cultural sexism using a composite index of items measuring aggregated social norms, consistent with numerous studies that have used similar attitudinal indices to measure various forms of macro-level stigma (e.g.,
Evans-Lacko et al., 2012; Hatzenbuehler et al., 2017b; Miller et al., 2011; Leitner et al., 2016) as well as cultural sexism, specifically (e.g., Charles et al., 2018). These items were obtained from two publicly available sources: (1) Project Implicit, which provided implicit gender attitudes from the Implicit Association Test (IAT) and (2) the General Social Survey, which provided explicit measures of gender role attitudes (e.g., whether it is better for women to stay home and care for children) and sexist beliefs (e.g., whether men are more emotionally suited for politics than women).

Based on theories of gender as a social structure (Risman et al., 2004; 2017), the work of Reid and colleagues (2014), and the emerging research on the mental health consequences of macro-level stigma (Hatzenbuehler, 2016), we hypothesized that therapies conducted in communities with higher (vs. lower) levels of cultural sexism would be less effective in study samples with a majority of girls. We examined whether this relationship remained robust after controlling for contextual factors that may co-vary with cultural sexism and intervention efficacy, and thus serve as confounders. We also explored whether study characteristics known to moderate the effectiveness of youth psychotherapy interventions (Weisz et al., 2017) may bias results if they were not randomly distributed across low- and high-cultural sexism studies in our sample. Finally, to determine whether this association was specific to samples with a majority of girls, a specificity analysis examined whether cultural sexism moderated intervention efficacy among study samples with mostly boys. The inclusion of this negative control analysis (Lipsitch et al., 2010) helps to improve inferences; if the association between cultural sexism and intervention efficacy is found only among majority-girl samples, it improves the likelihood that results are not confounded by other contextual factors associated with cultural sexism.

**Methods**
Information Sources and Study Selection

We examined a subset of studies from a larger database used in previous meta-analyses, described in detail elsewhere (Weisz et al., 2017, 2019). The larger database included peer-reviewed randomized controlled trials (RCTs) of youth psychotherapy published in English and identified in PsycINFO and PubMed between the years of 1963 and 2017. Studies were included if: (a) participants were selected and treated for depression, anxiety, conduct problems, and/or ADHD, (b) the mean sample age was between 4 and 18 years, (c) participants were randomly assigned to a treatment versus control condition and at least one treatment condition was psychotherapy, and (d) outcome measures were administered to both treatment and control groups. Studies were reliably coded for several study and sample characteristics, study quality indicators, and moderators examined in the original meta-analysis (additional data collection details available in Weisz et al., 2017, 2019).

Inclusion Criteria

In the present meta-analysis, studies were included if they (a) provided data on sample gender composition, (b) were conducted in the U.S. and (c) provided post-treatment scores on an outcome targeted by the intervention (e.g., a depression measure for a depression treatment RCT). With respect to the first inclusion criterion, nine studies occurred in multiple states. In sensitivity analyses, we removed these to ensure that their inclusion did not influence the direction or magnitude of the findings (see Online Supplement, Appendix S1). With respect to the third inclusion criterion, we examined clinical outcomes specifically targeted by the intervention, rather than including the diverse array of non-targeted, non-mental health measures sometimes included in study measurement models (e.g., time spent traveling to treatment
sessions, marital relationship of parents), consistent with previous meta-analyses focused on clinical outcomes that the treatments were designed to impact (e.g., Weisz et al., 2019).

Primary analyses were limited to studies (N=95) with samples comprised of 50% or more girls (see Online Supplement, References S1). We chose this cutoff for several reasons: (1) consistent with clinical research trials in general (Phillips & Hamberg, 2016; Welch et al., 2017), most youth psychotherapy studies do not report sex-specific outcome data (e.g., data allowing for the calculation of separate effect sizes for girls and boys); (2) a prior study examining associations between macro-level racism and intervention efficacy among majority-Black samples used this same cut-off (Reid et al., 2014); (3) it ensures sufficient statistical power (i.e., a more stringent cut-off would reduce the sample size); and (4) it avoids potential selection bias if we were to examine our research question using only the very small number of studies that were conducted with 100% girls (n=12). Given that gender composition of the studies is a binary variable, and that some studies included 50/50 girls/boys, using ≥50% boys would result in redundancies across models. Consequently, the negative control analysis is presented using intervention samples with greater than 50% boys.

Moderator Calculation

State-level cultural sexism was calculated using a composite index of 11 items that were compiled via explicit and implicit attitudinal indicators (n=19 candidate indicators) from 2 sources: Project Implicit (pooled across years 2003-2018) and the General Social Survey (pooled across years 1974-2014). The explicit indicators directly queried gender role attitudes (e.g., whether it is better for women to stay home and care for children) and sexist beliefs (e.g., whether men are more emotionally suited for politics than women, whether the gender disparity in STEM positions is due to differences in work ethic), whereas the implicit indicators were
obtained through the Implicit Association Test (IAT) examining to what extent respondents associate gender with career and scientific domains. These items were pooled and averaged to the state level (see Online Supplement, Appendix S2 and Table S1 for more details).

A factor score was created for each state by using exploratory factor analysis with a factor loading cut off of 0.50 (see all scores in Online Supplement, Table S2). The analysis was performed using PROC FACTOR in SAS 9.4, with the prior communality estimate fixed at squared multiple correlations with all other variables and oblimin rotation. Cultural sexism was also treated categorically using the “factor” function in the R metafor package (Viechtbauer, 2010), allowing us to simultaneously compare minimum, median, and maximum levels (i.e., tertiles) of cultural sexism observed in the final sample (Figure 1). Across the 32 states included in our analytic sample, cultural sexism ranged from -1.91 to 1.89 ($M$=-0.45, $SD$=0.86). The majority of studies ($n=67, k=694$) were conducted in states with lower levels of cultural sexism, with only 26 studies ($k=375$) conducted in states with cultural sexism higher than the national average (i.e., above 0).

To calculate county-level sexism across 86 counties in the analytic sample, we retained 8 of the 11 indicators from the state-level models that were also available at the county level (described in the Online Supplement, Appendix S2) and ran a confirmatory factor analysis with these items; counties were then assigned factor scores. As in the state-level analyses, the majority of studies in the analytic sample took place in counties with aggregate sexism levels lower than 0 ($n=89, k=997$). Across the 86 counties included in our analytic sample, cultural sexism ranged from -1.45 to 0.27 ($M$=-0.8, $SD$=0.36).

Covariates
We considered 6 state-level factors that could theoretically serve as common causes of both cultural sexism and intervention efficacy but that had not been used as indicators of macro-level sexism in prior work (e.g., religiosity): population density; two demographic characteristics associated with gender norms (percent foreign born, percent non-Hispanic White); two indicators of area-level income (percent poverty, median household income); and an indicator of inequality (Gini coefficient). Only two (Gini coefficient and median household income) were significantly associated with both cultural sexism and intervention efficacy, and were thus included as covariates in all analyses. In the Online Supplement, data sources, descriptions, and years are shown in Appendix S3 and the correlation matrix in Table S3.

We additionally examined two study-level variables and one effect size-level variable that predicted intervention efficacy in the larger meta-analytic database (Weisz et al., 2019): 1) targeted problem (e.g., internalizing problems); 2) informant (e.g., youth-reported symptoms); and 3) control condition (e.g., waitlist, usual care). While these are not confounders (i.e., they could not plausibly cause state-level cultural sexism), if these study-level characteristics exhibit similar geographic heterogeneity to state-level cultural sexism, it could spuriously influence results. That is, if the majority of study-level characteristics were coincidentally clustered in high-sexism states, it could create an apparent association between cultural sexism and intervention efficacy that was due to chance distributional differences alone. To rule out this possibility, we examined whether these characteristics were related to either cultural sexism or intervention efficacy and found that they were not; thus, these variables were not controlled in our analysis.

**Statistical Analyses**
Cohen’s $d$ was calculated for each study outcome measure (i.e., effect size; ES), reflecting the standardized mean difference between treatment and control groups at the end of treatment. ES was calculated using both sampling variation for each ES (Level 1) and within-study variation (Level 2). Hedges small sample correction was then added to all ESs to produce an unbiased estimate of the population standardized mean difference ($g$) (Hedges & Olkin, 1985). To reduce bias, all analyses weighted ES by the inverse of the sampling variance (Hedges & Olkin, 1985), and the residual degrees of freedom was used to compute the denominator degrees of freedom for fixed effects. Weighting increases the efficiency of estimates by giving more weight to ESs with greater precision (i.e., smaller standard errors; Hedges et al., 2010). Weighted two-level random-effects models (Cheung, 2014) were conducted with the metafor package of R Version 3.5.1. Random effects models allow for a distribution of true ES, rather than assuming one true effect exists across studies (Lipsey & Wilson, 2001; see Online Supplement, Appendix S4 for analysis code).

All studies included multiple outcomes (i.e., ESs), thus violating the assumption of independent ESs. Due to the many shortcomings associated with choosing just one ES or averaging ESs (Cheung, 2014), all ESs measuring a problem targeted by the intervention were included. Of note, adjustments for ES-dependency are unnecessary for meta-analyses aimed at examining general patterns, and excluding such adjustments provides conservative estimates (Hedges, 2006). Moreover, our moderator, cultural sexism, was measured at the study level (rather than the outcome level), thus partially accounting for ES-dependency (i.e., all ESs in a given study are associated with the same value of cultural sexism). Nevertheless, we ran an additional sensitivity analysis that accounted for clustering within studies. Although this model significantly reduced statistical power, results were in the same direction and magnitude as the
two-level model (see Online Supplement, Appendix S5), indicating that dependency did not bias our results.

Our analysis proceeded in several steps. We first tested an interaction between state-level cultural sexism and gender composition in predicting intervention efficacy to determine whether the association between cultural sexism and ES differed for samples with more girls (vs. boys). Finding that it did, we tested our primary hypothesis through a stratified analysis, where the moderating role of state-level cultural sexism on intervention efficacy was examined through meta-regression separately for samples with 50% or more girls, and for samples with more than 50% boys.

We ran two additional analyses to strengthen inferences. First, we used a more stringent cut-off to examine the association between state-level cultural sexism and intervention efficacy by restricting analyses to samples with 75% or more girls. Although this analysis reduced our statistical power, it enabled us to determine whether the association between state-level cultural sexism and intervention efficacy became stronger as the proportion of girls in the study sample increased. By increasing the specificity of our sampling criterion (i.e., increasing the threshold for majority girls from 50% to 75%), we would anticipate that the magnitude of the effect of cultural sexism should increase (i.e., reduce any bias towards the null), if the ES is indeed associated with sexism and not some related contextual factor. Second, while these previous analyses examined state-level cultural sexism as a continuous measure, we conducted an additional analysis examining this variable as a categorical moderator representing tertiles of cultural sexism (high, medium, low). This enabled us to determine whether we had evidence of a dose-response association between state-level cultural sexism and intervention efficacy among samples with 50% or more girls. Both of these approaches were designed to strengthen our
confidence in the validity of our main findings, as other contextual, state-level characteristics related to cultural sexism would be unlikely to exhibit either a higher magnitude of effects in a more specific subsample (i.e., super-majority girls) or a dose-response relationship.

Next, we examined associations between cultural sexism and intervention efficacy at the county level in samples with 50% or more girls. Finally, to examine the cross-level interaction between county- and state-level sexism, ESs were grouped into 4 categories based on whether the cultural sexism score was above or below the sample median: 1) high county/high state, 2) low county/high state, 3) high county/low state, 4) low county/low state.

Consistent with recommended practices for spatial meta-analyses, we also examined whether residuals in our primary model were correlated geographically (i.e., spatial autocorrelation) by calculating Moran’s I (Johnson et al., 2017). The Moran’s I test did not detect residual spatial autocorrelation, suggesting that model errors were randomly distributed in space (i.e., across states) and that no model adjustments were necessary.

**Results**

**Study Selection and Study Characteristics**

Of the larger database (Weisz et al., 2019), 4,233 ESs from 319 RCTs (N=20,513) met inclusion criteria for the current study (Figure S1, Online Supplement). These studies were published between 1963 and 2017. Our primary analytic sample included studies with samples of 50% or more girls (1,069 ESs; 95 RCTs, N=7,696). Therapies targeted a variety of problems, including ADHD (1 study), conduct (14 studies), anxiety (46 studies), depression (33 studies), and anxiety and depression (1 study). Studies tested one or more: youth-focused behavioral treatments (n=52), youth-focused non-behavioral treatments (n=22), caregiver/family-focused behavioral treatments (n=7), caregiver/family-focused nonbehavioral treatments (n=4), or
multiple types of treatments (e.g., youth-focused behavioral and youth-focused nonbehavioral treatments; \( n=10 \)).

Previously conducted tests of publication bias (i.e., a funnel plot, Egger’s weighted regression test, trim-and-fill, and an analysis of “zero-effect” studies) and study quality (i.e., via mixed models, and included measures of participant blindness to assessment, attrition, and measurement objectivity) indicated that publication bias was present but minimally impacted results, and that study quality was unrelated to ES (see Weisz et al., 2017, 2019 for details).

**Associations between Cultural Sexism and Intervention Efficacy**

A two-level random effects meta-regression analysis indicated a significant interaction between cultural sexism and gender composition \( (p<.0001) \). Stratified analyses showed that higher state-level cultural sexism was associated with significantly lower effect sizes for studies with \( \geq 50\% \) girls (adjusted \( \beta=-0.050, \) CI: -0.09, -0.01, \( p<0.05; \) Table 1); in contrast, state-level cultural sexism was unrelated to effect sizes in studies with \( >50\% \) boys \( (N=224, k=3164; \) \( \beta=0.002, \) CI: -0.02, 0.02, \( p=0.82) \). This result indicates that the association between cultural sexism and intervention efficacy was specific to samples with a majority of girls (Figure 2). The robustness of these results was further supported by an additional analysis of studies with 75% or more girls \( (n=26; k=357) \). The beta estimate was stronger than in studies with 50% or more girls (adjusted \( \beta=-0.12, \) CI: -0.22, -0.03, \( p<0.05) \), demonstrating that the association between state-level cultural sexism and smaller intervention effects became more robust as the proportion of girls in the study sample increased.

State-level cultural sexism was subsequently examined as a categorical moderator representing tertiles, with the values dummy coded and compared, to further probe this finding. Among intervention samples with 50% or more girls, ESs were lowest in states in the highest
tertile of state-level cultural sexism ($g=0.19$), compared to states in the middle ($g=0.21$) and lowest ($g=0.30$) tertiles (Table 1). ESs differed significantly between states with low and high cultural sexism ($t=-2.21, p<0.05$), but not between states with median and low cultural sexism, nor between median and high cultural sexism.

The beta estimate for county-level cultural sexism was slightly stronger than the state-level analysis, but did not reach statistical significance ($\beta=-0.065$, CI: -0.17, 0.04, $p=0.24$). However, there was a cross-level interaction between state- and county-level cultural sexism, which revealed that intervention studies with 50% or more girls had the strongest effect sizes when they were conducted in counties and states with the lowest levels of cultural sexism (predicted $g=0.34$, $p<0.001$). This predicted ES was significantly larger than the ESs for studies in the high county/high state group (predicted $g=0.24$, $t=2.1$, $p<0.05$), and in the other two groups (i.e., studies in low county/high state and in high county/low state; $ts=-3.3$ and -3.9, respectively, $ps<0.01$).

**Discussion**

Our spatial meta-analysis indicated that psychotherapy randomized controlled trials with samples comprised of a majority of girls were significantly less effective in states with higher vs. lower levels of cultural sexism. This association was observed even after control for contextual-level factors associated with cultural sexism and intervention efficacy, including another indicator of macro-level inequality (i.e., Gini coefficient). Importantly, the association between state-level cultural sexism and intervention efficacy became stronger as the proportion of girls in the study sample increased. Results also revealed an apparent dose-response relationship across categorical levels of cultural sexism, with substantive differences between the highest (predicted $g=0.30$) and lowest (predicted $g=0.19$) tertiles of state-level cultural sexism.
The estimate for county-level cultural sexism was comparable to that observed at the state level but did not reach statistical significance. This may be due to less precision at the county level, as fewer respondents contributed to the attitudinal measures at this geographic level, introducing measurement error in estimating the true mean values of county-level sexism. In addition, the range of cultural sexism at the county level was smaller than that at the state level, restricting variability, and thus statistical power. However, the cross-level interaction between state- and county-level cultural sexism was statistically reliable, indicating that the strongest benefits of psychotherapy studies with a majority of girls were observed in states and counties with the lowest levels of cultural sexism. Finally, the fact that state-level cultural sexism was unrelated to intervention efficacy for samples with a majority of boys suggests that our results are specifically attributable to cultural sexism, rather than to associated factors. In other words, if cultural sexism is merely a proxy for other contextual factors (e.g., socioeconomic status, conservatism), it should operate similarly (i.e., be associated with reduced treatment efficacy) among samples of both boys and girls.

Spatial meta-analysis is uniquely suited to addressing our research question, because it capitalizes on the substantial heterogeneity in exposure to cultural sexism that occurs across individual studies. Because these individual studies are typically conducted in a single location, the context is invariant, and thus it is not possible to detect associations between contextual factors and intervention efficacy. As such, these results provide a novel demonstration that contextual features related to cultural sexism may undermine the efficacy of psychotherapy interventions comprised of samples with a majority of girls.

At the same time, our approach is less well-suited for answering questions of mechanism—that is, why therapies in high cultural sexism states are less effective among study
samples with more girls. Given established associations between community-level prejudice and adverse outcomes (Hatzenbuehler, 2016; 2017a, b), it is possible that girls in communities with high levels of cultural sexism arrive at treatment with elevated mental health symptoms and other psychosocial problems that render them less able to derive benefit from psychotherapy. Alternatively, girls may enter treatment in high and low sexism communities with similar levels of psychopathology, but cultural sexism may increase risk for certain psychosocial processes that negatively affect treatment engagement and/or retention once therapy is initiated. In preliminary support of this idea, macro-level stigma is associated with higher levels of social isolation (Perales & Todd, 2018), hypervigilance and hopelessness (Russell & Richards, 2003), and perceived stress (Rostosky et al., 2009), some of which impede upon intervention uptake or response, or both (Alfano et al., 2009; Gallagher & Resick, 2012). Finally, it is possible that cultural sexism continually undermines any gains made in psychotherapy because these contexts lack the type of structural supports necessary for girls to incorporate the skills learned in therapy into their daily lives. We are unable to test these competing explanations because, like all meta-analyses, we are limited by the data that could be reliably coded across all of the studies included in the meta-analytic database. Future research is therefore needed to identify the causal mechanisms underlying our findings, which can point to the most effective ways to modify psychotherapies delivered in environments characterized by high levels of cultural sexism to ensure that they are maximally efficacious.

Our results are consistent with those of Reid and colleagues (2014), who found that a health behavior intervention for African Americans was less effective in areas high in racial prejudice. Thus, there is now evidence across two studies focusing on different forms of macro-level stigma and prejudice (race, gender) and different intervention foci (HIV prevention
interventions, psychotherapy) showing that psychological interventions conducted in
communities with higher levels of stigma and prejudice are less effective than those conducted in
low-stigma contexts. It is therefore worth considering the implications of these emerging results
for clinical and public health interventions. If future research continues to replicate this finding,
there would be additional evidence that cultural sexism should be addressed via structural- and
community-level interventions to enhance the success of psychological interventions conducted
and disseminated in these contexts. While structural interventions to reduce sexism are rare,
examples of structural approaches in other areas exist (e.g., Chaudoir et al., 2017), supporting the
utility of this approach.

At the same time, changing social structures can be a protracted process, and thus clinical
interventions are necessary to assist individuals in high-sexism contexts to cope effectively. In
their meta-analytic study, Reid et al. (2014) found that interventions that tailored content to
participants’ values and needs buffered against the adverse impact of structural racism on
African Americans’ condom use, perhaps by reducing respondents’ distrust of intervention
providers. The extent to which analogous forms of tailoring may improve girls’ therapy response
in the context of cultural sexism remains an open question. Further, recent research (Pachankis et
al., 2020) has demonstrated that brief, online interventions that promote personal (e.g., self-
acceptance) and interpersonal (e.g., social support) coping are effective in reducing
psychological distress among sexual minority youth living in high-stigma, low-resource settings
(i.e., Appalachia). This dual focus on individual and structural levels aligns with best practices in
intervention research, as it is increasingly recognized that multi-level approaches are needed to
reduce stigma and its negative consequences (e.g., Rao et al., 2019).
Our study has several limitations. First, our sample included only 32 states and 86 counties, which resulted in a restricted range of our cultural sexism variables that over-represented treatment trials in states and counties with lower levels of cultural sexism (see Online Supplement, Appendix S1). However, this restricted range would have reduced our statistical power to detect an association; thus, we have likely underestimated the association between cultural sexism and psychotherapy efficacy among samples with a majority of girls. In addition, this finding is important in its own right, because it suggests that psychotherapy interventions may not be currently disseminated in the communities where they are most needed (i.e., in high cultural sexism states and counties).

Second, the datasets that we used to create the cultural sexism index had small sample sizes of survey respondents in most states and counties within individual years. Consequently, cultural sexism was aggregated across years to create a more stable estimate, and then analyzed as a time-invariant predictor. One potential limitation of this approach is that it does not capture changes in temporal trends in sexism. However, previous analyses of attitudinal data from the General Social Survey showed that while sexism decreased over time, the ranking of each state’s measure of cultural sexism relative to other states was stable across time (Charles et al., 2018). This finding indicates that the places with higher cultural sexism in previous decades remain so today despite national declines in sexism overall, suggesting that a time-invariant measure represents a valid measure of cultural sexism. Nevertheless, future research should determine whether the results obtained here are consistent when modeling cultural sexism as a time-varying moderator of intervention efficacy.

Third, while the concept of macro-level sexism—in both its material and cultural dimensions—has existed for decades (e.g., Epstein, 1988; Kanter, 1977; Risman, 2004), its
operationalization in the health literature is a relatively recent occurrence, and most of the research has focused on examining material forms of macro-level sexism (e.g., Homan, 2019). We advance this literature by measuring cultural dimensions of macro-level sexism through the use of state- and county-level attitudinal measures. However, given that this field is still in its infancy, there is not yet a consensus on the best way to measure this construct. While we offer an empirically-derived approach to quantifying cultural sexism, further research is needed to determine the relative validity and precision of competing measurement approaches. In addition, future research could examine whether our results are generalizable to material dimensions of macro-level sexism, and whether material or cultural dimensions are more strongly associated with intervention efficacy among girls.

This is, to our knowledge, the first study to document that the social context surrounding girls may partially explain heterogeneity of treatment effects in therapies. Our findings suggest new avenues of research investigating whether existing psychotherapies might be adapted to enhance their effects among girls in environments with pronounced cultural sexism, and possibly a need for structural- and community-level interventions to target cultural sexism at its source. This research also joins an emerging body of evidence suggesting that the efficacy of psychological interventions may depend in part on the social contexts in which they are delivered (Johnson et al., 2017; Reid et al., 2014; Yeager et al., 2019).
References


Leitner, J. B., Hehman, E., Ayduk, O., & Mendoza-Denton, R. (2016). Blacks’ Death Rate Due to Circulatory Diseases Is Positively Related to Whites’ Explicit Racial Bias: A


Figure 1. Observed Cultural Sexism in Sample of Studies with Majority Girls (N = 95)
Figure 2. Cultural Sexism and Effect Size Across Subsets by Majority Sex
### Table 1
Psychotherapy effect sizes as a function of cultural sexism

<table>
<thead>
<tr>
<th>Variable</th>
<th>β</th>
<th>SE</th>
<th>95% CI</th>
<th>Moderator t-test (t-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Unadjusted model:</strong> Cultural sexism without covariates</td>
<td>-0.06</td>
<td>0.02</td>
<td>[-0.10, -0.01]</td>
<td>-2.55*</td>
</tr>
<tr>
<td><strong>Adjusted model:</strong> Cultural sexism controlling for contextual covariates</td>
<td>-0.05</td>
<td>0.02</td>
<td>[-0.09, -0.01]</td>
<td>-2.33*</td>
</tr>
<tr>
<td>Gini coefficient</td>
<td>0.02</td>
<td>0.01</td>
<td>[-0.002, 0.04]</td>
<td>1.75</td>
</tr>
<tr>
<td>Median household income</td>
<td>0.03</td>
<td>0.02</td>
<td>[-0.01, 0.07]</td>
<td>1.71</td>
</tr>
</tbody>
</table>

**Tertile Analysis**

<table>
<thead>
<tr>
<th>Cultural sexism [range]</th>
<th>k</th>
<th>Predicted g</th>
<th>SE</th>
<th>95% CI</th>
<th>Moderator t-test (t-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum [-1.91, -0.87]</td>
<td>377</td>
<td>0.30</td>
<td>0.03</td>
<td>[.24, .36]</td>
<td>9.38***</td>
</tr>
<tr>
<td>Median [-0.86, 0.019]</td>
<td>394</td>
<td>0.21</td>
<td>0.03</td>
<td>[.15, .28]</td>
<td>6.52***</td>
</tr>
<tr>
<td>Maximum [0.19, 1.90]</td>
<td>298</td>
<td>0.19</td>
<td>0.04</td>
<td>[.12, .26]</td>
<td>5.14***</td>
</tr>
</tbody>
</table>

*Note:* Predicted effect sizes (predicted g) are given at the minimum, median, and maximum values observed in the data. *k* = number of effect sizes in each category. Results were produced using (1) random-effects meta-regression models excluding the intercept and including the "factor" function in the R metafor package, and (2) Knapp and Hartung adjusted t-tests (Viechtbauer, 2010). A significant moderator t-test result indicates that the effect for the cultural sexism tertile level is significant after adjusting for the other levels. Gini coefficient is a summary index of income inequality ranging from 0 (perfect equality) to 100 (perfect inequality); it was available from the US Census American Community Survey 5-year estimates, which were pooled and averaged for the years 2006-2010. Median household income was available from the US Census Current Population Survey and was pooled and averaged across all years 2000-2010.

*p < .05, **p < .01, ***p < .001