REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
16/10/2020	Final Report	17/06/2019 - 16/07/2020

4. TITLE AND SUBTITLE
Final Report: Theore: Theory-Driven Curation and Robustness Calculus of Social and Behavioral Sciences (SBS) Claims

5a. CONTRACT NUMBER
N/A

5b. GRANT NUMBER
HR00111920023

5c. PROGRAM ELEMENT NUMBER
N/A

6. AUTHOR(S)
Nan Zhang
Heng Xu

5d. PROJECT NUMBER
N/A

5e. TASK NUMBER
N/A

5f. WORK UNIT NUMBER
N/A

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Amerian University
4400 Massachusetts Ave NW
Washington, DC 20016

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency
675 N Randolph St, Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
DARPA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
N/A

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
N/A

14. ABSTRACT
The goal of our Theore project is to demonstrate the feasibility of computationally reasoning about the robustness of SBS claims based on a quantitative model that captures the connections between SBS claims, including and especially conflicting ones, through their shared constructs, relationships, theories and studies. In this project, we developed novel mixture-decomposition algorithms specifically for Theore to meet this goal.

15. SUBJECT TERMS
robustness of scientific claims; meta-analysis; mixture decomposition

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Nan Zhang
N/A	N/A	N/A	N/A	16	**19b. TELEPHONE NUMBER** *(Include area code)* (202) 885-1765

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATE COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33315-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report. e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

Final Report

Project Number: HR00111920023

Project Title: *Theore*: Theory-Driven Curation and Reusable Robustness Calculus of Social and Behavioral Sciences (SBS) Claims

Principal Investigators: Nan Zhang (nzhang@american.edu) and Heng Xu (xu@american.edu)

Performance Period: June 17, 2019 - July 16, 2020

1. Project Objectives and Goals

The very essence of scientific progress is the systematic accumulation of knowledge, yet the viability of doing so is being questioned in social and behavioral sciences (SBS) by experts and practitioners alike. Critics point to many conflicting findings on the same research claim, akin to a famous quote by Senator Walter Mondale: "*For every study that contains a recommendation, there is another, equally well documented study, challenging the conclusions of the first... No one seems to agree with anyone else's approach*". Without the ability to collectively reason about such conflicting findings, we often see tens, even hundreds, of studies examining a research claim from different angles, only to make the picture even murkier for a practitioner who wants a firm answer.

The goal of our *Theore* project is to demonstrate the feasibility of computationally reasoning about the robustness of SBS claims based on a quantitative model that captures the connections between SBS claims, including and especially conflicting ones, through their shared constructs, relationships, theories and studies. In this project, we developed novel mixture-decomposition algorithms specifically for *Theore* to meet this goal.

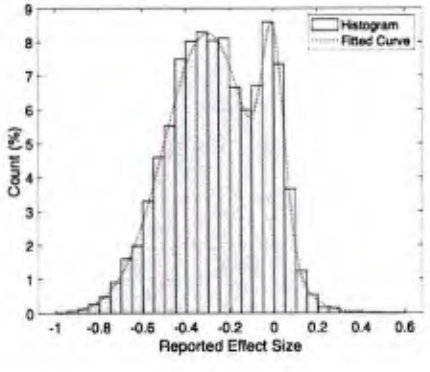
2. *Theore* Design Rationale

2.1. Overview of Mixture-based Heterogeneity Analysis

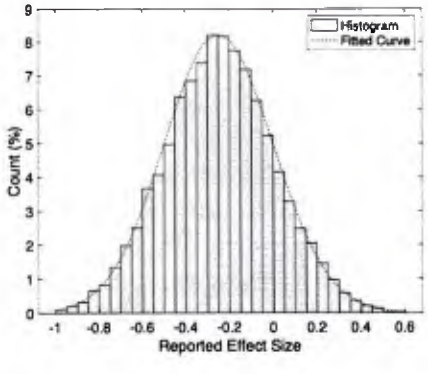
To address the "heterogeneity" problem stated in Senator Mondale's famous quote, scientists often conduct a "meta-analysis" to statistically combine the results of multiple studies to identify patterns among different or even conflicting findings. Such investigations have led to scientific breakthroughs in many disciplines, from psychology (Huedo-Medina et al, 2006) to epidemiology (Berlin, 1995) to medicine (Higgins et al, 2003). Going through these success cases, a commonality emerges: New discoveries often ensue once a meta-analysis researcher successfully *disentangles* the various subsets of primary studies that caused the observed heterogeneity. For example, in a classic meta-analysis of the effect of oral-contraceptive use on breast cancer, Romieu et al (1990) demonstrated that the heterogeneity of primary studies is mostly attributable to the difference between two subsets -- those studying premenopausal subjects and those also involving postmenopausal subjects. By disentangling the two components, Romieu et al were able to isolate the hypothesized effect to the first subset -- a critical insight that was later confirmed by other studies (Berlin, 1995).

Despite the importance and value of disentangling heterogeneity, actually doing so in a meta-analysis can be quite challenging, not only because it tends to require significant manual efforts, but the procedure of doing so varies considerably even within a field (Fletcher, 2007; Naaktgeboren et al., 2014). To start, whether the observed yet unexplained heterogeneity is "natural" or warrants further investigation is a matter of subjective judgment (Higgins et al, 2003). In exploring sources of heterogeneity, researchers often hypothesize moderator variables[1]. Not only does coding such variables require substantial resources, but the difficulty, and hence precision, of coding often varies significantly from one primary study to another. Furthermore, it is often challenging to determine whether coding finer gradations for the moderator variable could resolve the remaining heterogeneity (Hunter & Schmidt, 2004, p. 180); or if there are other unknown artifacts causing it. Last but not least, when heterogeneity is partially caused by availability biases such as publication bias or questionable research practices (QRPs) (e.g., *p*-hacking, Simonsohn et al, 2014), the existing statistical tools often have difficulty disentangling the heterogeneity of "ground-truth" effects and that caused by availability biases (McShane et al, 2016). Hence, these tools struggle to answer simple questions like how much percentage of primary studies are affected by such biases, or what meaningful knowledge we can retain about an effect if the existing studies of it collectively exhibit some traces of QRPs.

In summary, while many methods[2] are available to support the *deductive* approach of hypothesizing and then testing a specific source of heterogeneity, we lack an *inductive* "data-driven" method that can infer the nature of heterogeneity and its likely causes directly from the reported outcomes of primary studies. Consider a simple example in Figure 1: While the reported effect sizes in both cases feature the exact same mean and variance, even a casual inspection of their distributions would reveal an intriguing pattern in Case A: those primary studies reporting an effect size between -0.1 and 0.1 form a distinct subgroup that shows little heterogeneity among themselves, yet is obviously different from the rest of the studies. Comparing the two cases side-by-side, it is easy to see that the heterogeneity in Case A is much less likely to be "natural" than in Case B -- a conclusion that cannot be drawn from the mean and variance of the effect sizes alone.



Case A: The reported effect sizes form a bimodal, skewed, distribution.



Case B: The reported effect sizes form a unimodal, largely symmetric, distribution.

Figure 1. Motivating examples for a data-driven inductive approach

[1] A moderator variable changes the strength or direction of an effect between two *variables* x and y. In other words, it affects the relationship between the independent *variable* or predictor *variable* and a dependent *variable*.

[2] For example, meta-regression for moderator variables, funnel plot for publication bias.

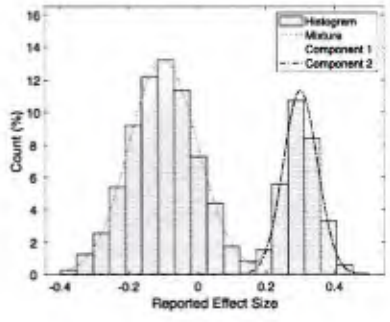
We develop a novel meta-analytic approach that *lets the data (reported by the primary studies) reveal the nature of the heterogeneity and the potential causes of it.* A key premise of our approach is a statistical theory known since the 19th century: if there is heterogeneity in the underlying data-generation processes (i.e., the ground-truth effect sizes captured by primary studies), then the distribution of the observed data (i.e., the effect sizes reported by primary studies) is bound to be a mixture of multiple distributions (Böhning, 1999). Such a mixture distribution would have samples drawn with probability w_i from its i-th component distribution ($i \in [1, m]$), where $w_1 + \ldots + w_m = 1$.

A historically famous example of mixture distribution is the Pearson's crab data: After collecting biometrics from a sample of 1,000 crabs at Naples in 1893, zoologists found the distribution of a key metric, the "forehead"-breadth-to-body-length ratio, significantly deviates from a Gaussian distribution. Pearson (1894), upon analyzing the data, found that the distribution of the metric closely resembles a mixture of two Gaussian distributions (with respective weights of 41% and 59%), indicating that the sampled crabs came from not one, but two different species. Besides Pearson's crab data, numerous mixture distributions have been identified in a variety of domains, from adult heights in the US (Kalai et al, 2016) to pricing strategies in online marketplaces (Seim & Sinkinson, 2016). One can see from these examples that the underlying heterogeneity (e.g., crab species) caused the observed data to be a mixture of multiple distributions -- a distribution that is noticeably different from a single Gaussian. Such a difference would in turn allow a researcher to decompose the mixture and identify the underlying heterogeneity, like what Pearson did to reveal the two different species in the crab samples. In the same spirit, if there is significant heterogeneity among primary studies, then their reported effect sizes should also form a mixture distribution which, upon careful examination, could be decomposed into heterogeneous components to reveal their distinguishing factors. In a nutshell, our approach draws from recent breakthroughs in theoretical machine learning (Belkin & Sinha, 2010; Kalai et al., 2010; Moitra & Valiant, 2010) to automate the decomposition process, thereby revealing the nature and likely sources of heterogeneity.

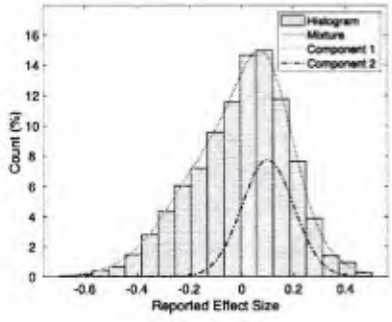
Interestingly, meta-analysis researchers have frequently noticed that the distribution of reported effect sizes has little resemblance with a Gaussian distribution (Micceri, 1989), but is instead closer to a *mixture* of multiple distributions: For example, Chernev et al (2015) observed in the meta-analysis for a controversial problem that most primary studies were "designed to document" how the direction of an effect can be reversed by adjusting a moderator variable, making it likely for their reported effect sizes to be a mixture of two distributions, one with a positive mean and the other negative (like Figure 2a). Hunter & Schmidt (2014) also recognized that, when the distribution of a moderator variable is a continuum in primary studies, the observed distribution of effect sizes could be a mixture of numerous distributions, each with a different mean and variance. Yet others hypothesized that the effect sizes reported by "outlier" primary studies from a distribution with a larger variance than the rest of the studies (Beath, 2014), making the overall distribution a mixture of both.

While recognizing the presence of mixture distributions is nothing new, what is novel in our *Theore* development is the idea of using an automated method to *decompose* a mixture distribution into likely components that explain the underlying heterogeneity. Interestingly, despite the wide recognition of mixture distributions in meta-analysis and the extensive statistics literature on Gaussian mixture models (McLachlan & Basford, 1988), the analysis of mixture models in meta-analyses was exceedingly rare, with only a few notable exceptions in medicine-related fields for limited purposes such as outlier detection (Schlattmann et al, 2015; Nord et al, 2017). One likely reason for the absence is the computational challenges associated with separating distributions that *overlap* with each other (see Figure 2b) -- a

challenge exemplified in meta-analysis given the limited number of primary studies and the oft-small or medium effect sizes studied in fields like management (Paterson, et al., 2016) and social psychology (Richard et al, 2003). To understand why overlapping components pose a significant challenge to mixture decomposition, consider Figure 2: If the two component distributions are well separated from each other so as to form two separate "peaks" in the mixture (Figure 2a), then decomposing the mixture into the two components is trivial so long as there are enough samples (i.e., primary studies) to reveal the peaks. But if the two distributions largely overlap with each other so as to form only a single peak (Figure 2b), the decomposition becomes much less obvious.



(a) an easy-to-decompose mixture. Note that the mixture curve largely overlaps with the two components because of the significant separation between them



(b) a hard-to-decompose mixture. Note the significant overlap between the two components.

Figure 2. Examples of Gaussian mixtures

Traditionally, the studies of mixture decomposition in statistics and computer science confirmed this intuition. Before 2010, researchers mostly followed a clustering-based method (e.g., Dasgupta, 1999) which starts with inferring which component each sample belongs to before computing the summary statistics of samples from the same component. This method fails when the components overlap significantly, because it is impossible to accurately determine for a sample in the overlapping region (e.g., an effect size in [0, 0.2] in Figure 2b) which component it belongs to. A trio of breakthroughs in 2010 in theoretical machine learning (Belkin & Sinha, 2010; Kalai et al., 2010; Moitra & Valiant, 2010) solved this problem by directly inferring the mixture structure, i.e., the parameters and weight of each component, without attempting to determine the component membership of each sample first. By doing so, the state-of-the-art algorithms can successfully decompose a mixture even when the components overlap almost entirely with each other (Kalai et al, 2016).

Drawing from this technical breakthrough, we develop novel mixture-decomposition algorithms specifically for *Theore* to run more advanced meta-analysis for disentangling the observed heterogeneity across studies.

2.2. *Theore* Design Features

In this section, we discuss the major design features of *Theore*, and how the output of mixture decomposition generated by *Theore* can be used in many different ways to enrich our understanding of the focal effect in a meta-analysis.

Understanding the Robustness of an Effect: Practitioners who want to leverage an effect in practical settings often want to have a robust, perhaps even conservative, estimate of the effect size. Especially at

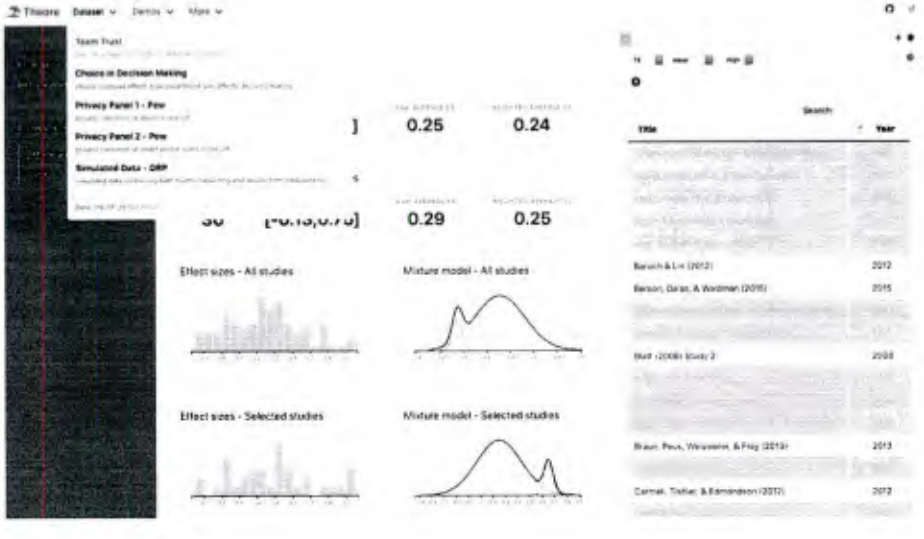
a time when many QRPs are identified and discussed, it becomes critical for a meta-analysis to offer practitioners the reassurance that, even after taking into account the possible presence of QRPs, the estimated effect size could still justify its implementation in practice. The mixture decomposition function of *Theore* supplies an effective tool to provide such an assurance, as the output component with the smallest effect size naturally becomes a conservative estimate. No matter if the other component is caused by QRP or other study-level characteristics, a practitioner can always inspect the more conservative component to determine whether it is worthwhile to leverage the effect in practice.

Heterogeneity Analysis: An important goal of meta-analysis to determine how well the observed heterogeneity can be accounted for by known factors such as sampling error, measurement artifacts, moderator variables, etc. (Hunter & Schmidt, 2004). Nonetheless, existing meta-analysis methods usually stop at this identification task. In other words, once a meta-analysis finds substantial heterogeneity that is left unexplained, it often has little to offer on finding an explanation for the remaining heterogeneity, other than simply suggesting researchers to continue studying moderator variables that may do so. While the output of mixture decomposition also cannot automatically explain the remaining heterogeneity, *Theore* does provide exploratory support that ease this manual task.

Supplementing or Corroborating Manual Coding: The study of moderator variables in many existing meta-analyses rely on the manual coding of these variables for each primary study. While numerous precautions have been suggested to increase the reliability of manual coding -- from asking for inputs from the original authors to intercoder reliability checks -- the difficulty of coding, hence the accuracy of coded variables, inevitably varies across different primary studies. Furthermore, there are study-level variables that cannot be properly coded even with manual efforts -- an example is whether QRP was involved in a study. In all these cases, the output of mixture decomposition can supplement or corroborate the results of manual coding. *Theore* can verify the results of manual coding, or selectively code certain studies in the non-overlapping region when the corresponding manual coding is infeasible.

2.3. *Theore* Implementation and User Interface

We now briefly discuss the user interface of the *Theore* system. The following screenshot shows the overall system interface of *Theore*. It consists of four main components: the top menu bar, the left navigation panel, the right data exploration panel, and the middle data-analytics panel.

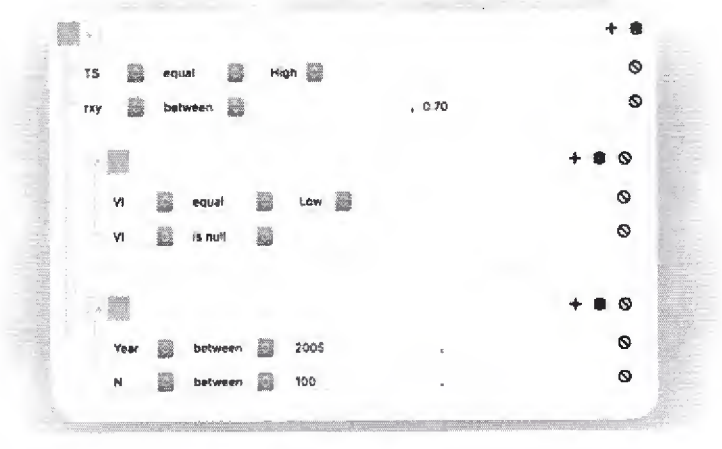


A key feature offered by the top menu bar is the ability to choose the dataset to analyze. Here a dataset could be a collection of results reported by primary studies, akin to the input to a meta-analysis, or a direct collection of records, like (shown in the figure as an example) the privacy-attitude panel data collected by the Pew research center. Besides the data itself, additional information specified for a dataset in *Theore* includes the focal attribute of analysis (e.g., which attribute represents the effect size reported by a primary study), the moderator variables of interest, etc. Besides the dataset-selection feature, the top menu bar also includes links to demonstrations (of how to use the *Theore* system), documentation, etc.

The right-side data exploration panel consists of two main components: the bottom table display and the top query-specification tool. By default, the display component shows all attributes of all records in the selected dataset. A user can order by any attribute, or perform a keyword search using the search box. A user can also select any arbitrary subset of data (using shift key + mouse clicks), such that the data analysis on the middle panel is limited to only the selected subset.

While the manual-select operation may be convenient for removing (or adding) one or a small number of studies (or records) from consideration, a more efficient way to select a subset of studies is to use the query-specification tool at the top of the data display panel. The following screenshot shows an example of using the query-specification tool. One can see from the screenshot that this tool allows the specification of any logical expressions involving equality or range conditions of any subset of attributes. The query specified in the screenshot is:

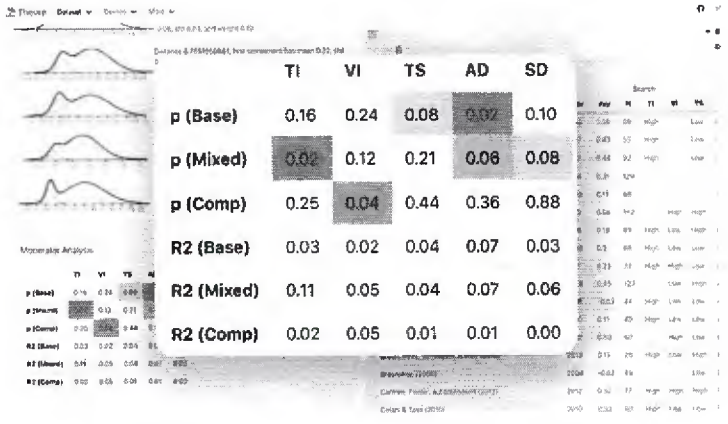
(TS = HIGH) AND (rxy < 0.70) AND ((VI = LOW) OR (VI IS NULL)) AND ((YEAR > 2005) OR (N > 100))



Here "NULL" is referring to cases where the attribute value is missing from the input data. Note that when one of the two inputs for a "between" condition is empty, the system only enforces the inequality condition with the specified input. An example is the conjunctive condition of rxy < 0.70 in the screenshot. Also note that, while the "between" condition does not include equality, it is easy to specify a less-than-or-equal-to condition like rxy ≤ 0.70 with (rxy < 0.70) OR (rxy = 0.70).

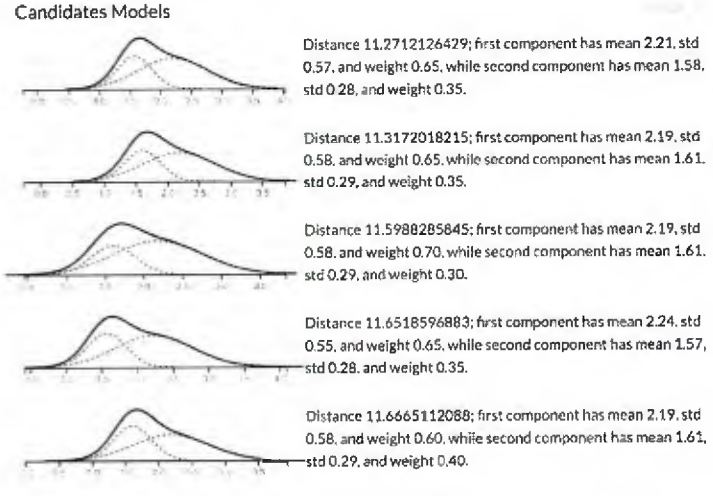
Once a user specifies a query with the query-specification tool, the studies satisfying the specified query condition will be automatically selected in the data display, and used as input to data analysis in the middle panel. A user can then perform minor adjustments to the selection by holding the shift key while using mouse clicks on the data display. Note that, while the query selection condition will not change in

response to the user adjustments, the input to the data analytics panel will change as the selection changes.

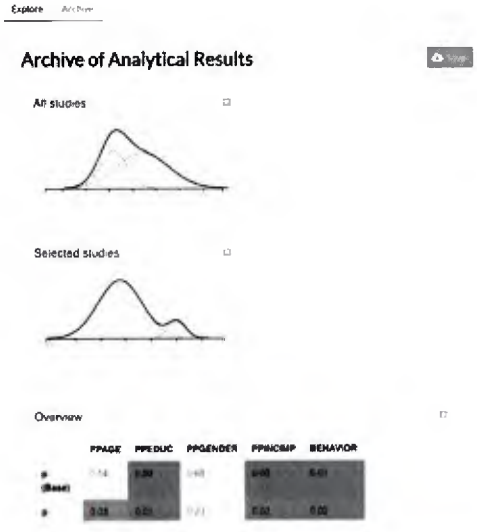


Finally, the middle panel displays the results of our mixture-based data analysis on the input data. The results of data analysis include multiple components: The top one consists of basic statistics such as the number of studies (or records), the mean and variance of raw and corrected effect sizes, etc. Following the basic statistics, the panel displays the histogram of the raw and corrected effect sizes, compares such histograms with those of all studies (instead of just the ones selected on the right-side data exploration panel), and displays the outcomes of mixture decomposition. After those, the panel displays various dedicated components for common tasks in a meta-analysis. For example, the above screenshot shows a component that uses the outputs of mixture decomposition to improve moderator analysis. It displays the p-value and R^2 when we use the reported effect size and the mixture component affinity as the regressant in meta-regression, respectively.

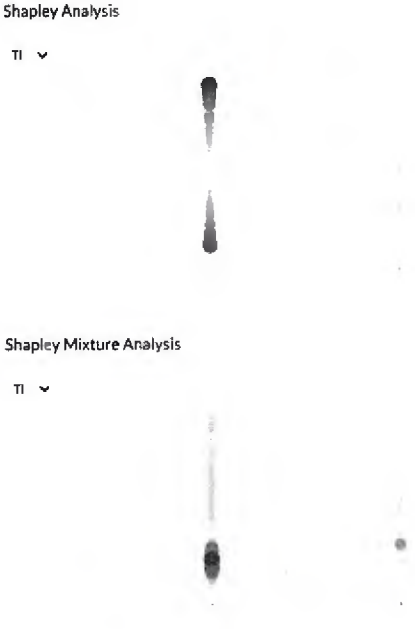
The following screenshot shows a component that displays the top candidates for the mixture composition. One can see that a user has several options on interacting with the component: The i icon allows a user to learn more information about the analytical results, including typical patterns that can be recognized from the visualization. The minimization icon allows a user to minimize the box if the result is not of interest. The camera icon allows a user to capture the results into the archive tab.



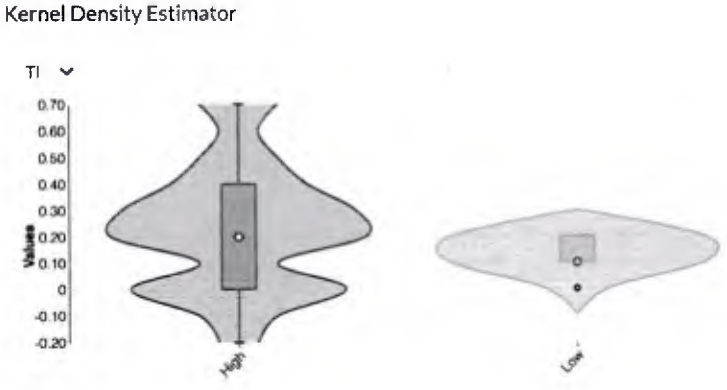
The following screenshots show the functionalities provided under the "archive" tab. It consists of multiple analytical results captured by the user (using the camera icon). For each result, it also captures the subset of data that was used to generate the analytical results. In addition, there is a "Save" button which exports the captured results (and data) to a DOC file for the user to download.



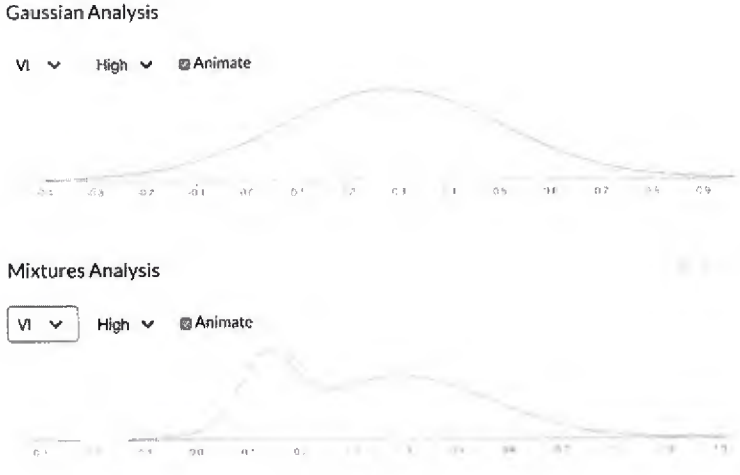
The following screenshots demonstrate a few new features introduced to the *Theore* system. First is a visualization of the importance of each study according to our Shapley-based analysis. Each circle corresponds to one primary study. The *x* axis is corresponding to any variable value (e.g., the moderator level for a hypothesized moderator variable), the *y* axis is corresponding to the reported effect size, and the radius of each circle is proportional to its "importance" on the output of the meta-analysis (e.g., the output of significance test for moderator analysis). As such, a primary study corresponding to a "larger" circle has more impact on the outcome of meta-analysis than a primary study corresponding to a smaller circle.



The following screenshot shows the visualization of a kernel-density estimator for the effect-size distribution, i.e., a non-parametric estimate of its probability density function.



We also added animations for contrasting the traditional model of a single Gaussian distribution with our mixture-based model, as shown in the following figure. One can see that such an animation clearly demonstrates how the finer resolution offered by our mixture-based model can boost the statistical power of meta-analytic tasks such as moderator analysis.



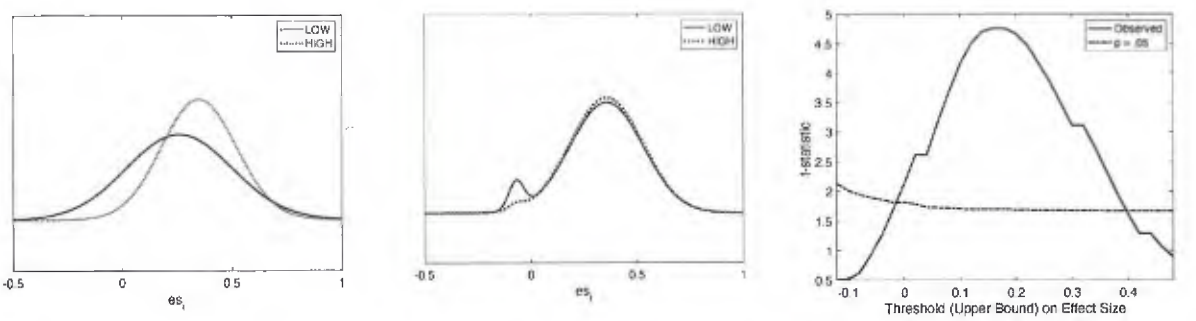
3. *Theore* Case Studies

3.1. Case Study: Using *Theore* to Disentangle Effect Size Heterogeneity

In this case study, we considered a problem extensively studied in social psychology: whether trust matters more for the performance of virtual teams than face-to-face teams (i.e., whether team virtuality has a moderating effect on the relationship between intra-team trust and team performance). Interestingly, the two recent meta-analyses, published on the same issue of the Journal of Applied Psychology, drew different conclusions: While Breuer, Hüffmeier, and Hertel (2016) found it significant (i.e., the trust-performance relationship was stronger in virtual teams than face-to-face teams), De Jong, Dirks, and Gillespie (2016) found that the strength of the trust-performance relationship does not meaningfully differ between virtual teams and face-to-face teams. These different conclusions continued

the long-standing inconsistency in the literature on whether the moderating role of team virtuality indeed exists.

Throughout this case study, we used the same data and artifact-correction procedures as De Jong et al. (2016). Specifically, we first downloaded the data from its supplemental materials, and then applied sampling-error correction and measurement-error correction according to the specifications within. We then applied *Theore*'s latent mixture-based method on the data and our results did identify a significant moderating effect of team virtuality. To understand how our latent mixture-based method was able to identify the moderating effect while the existing method cannot, we compared how traditional subgroup analysis and our method model the (subgroup) distributions for face-to-face teams (i.e., moderator = LOW) and virtual teams (i.e., HIGH). As can be seen in Figure 3b, our latent mixture-based model reveals that the two moderator levels differ quite significantly when the effect sizes are small, yet the difference diminishes when the effect sizes are large. Unfortunately, if one did not disentangle the mixture components but instead relied on the overall mean and standard deviation of the two subgroups, like in traditional subgroup analysis, this difference would no longer be recognizable, as shown in Figure 3a. To verify this explanation, we designed a simple test with the aim of "isolating" the effect of the right-most component. Specifically, we considered the following question: if we only considered primary studies that reported effect size below a certain threshold, say 0.2, we would have "zoomed in" to the part of the distribution where the moderator variable has a significant effect. In this case, would it be more likely for a moderator-analysis method to identify the moderating effect despite of the reduced sample size? As can be seen in Figure 3c, when the threshold increased, the t-score first rose until reaching the peak when the threshold was around 0.2, after which it declined, eventually falling out of the statistically significant range. This is remarkably consistent with our explanation that the increased resolution offered by our method is the reason why it detected the moderating effect.



3a) Models based on a single Gaussian distribution

3b) Models based on our mixture decomposition results

3c) Changes of t-statistic with threshold on reported effect size

Figure 3. An illustration of our case study for examining the moderating effect of team virtuality in the relationship between intrateam trust and team performance. Note - The left figure depicts the distributions being tested in traditional subgroup analysis. Since the outcome depends only on the mean ρ and standard deviation SD_ρ of the corrected effect sizes, the traditional method is essentially testing the mean difference between the two Gaussian distributions depicted in the figure (with mean ρ and standard deviation SD_ρ for each subgroup). The middle figure depicts what is tested in our method, i.e., the component affiliation. Each line represents the estimated mixture distribution for the corresponding subgroup, generated by weighting the decomposed mixture components with the posterior distribution of component affiliation for effect sizes in the subgroup. The right figure depicts the change of the aforementioned t-statistic (for the difference between the mean effect sizes of the two subgroups) when we only take as input the primary studies with reported effect sizes below a threshold, which varies from -0.1 to 0.5. The dotted line in the figure marks the t-score corresponding to $p = .05$.

In this case study, the latent mixture-based method not only identified the moderating effect, but also pinpointed where team virtuality likely had the strongest effect: when the effect size of the trust-

performance relationship was relatively small. There are various explanations for why this could have happened. For example, it could have been caused by interactions between team virtuality and other team characteristics, as pointed out by De Jong et al. (2016) and De Guinea et al. (2012). It might also have been caused by a "ceiling effect" on how strong the trust-performance relationship could be. More primary studies should be conducted to understand the reason behind this observation.

More details about this case study are available via our working paper, which is accepted by *Psychological Methods*:

Zhang, N., Wang, M., and Xu, H. (2020). "Disentangling Effect Size Heterogeneity in Meta-analysis: A Latent Mixture Approach," *Psychological Methods*, (in press).

3.2. Case Study: Using *Theore* to Detect Heterogeneity Caused by Non-Monotonic Effects

In this case study, we considered a problem extensively studied in consumer psychology: choice overload, which captures the idea that offering more options could impede rather than improve consumer satisfaction. Following the dramatic evidence of choice overload in social psychology, a series of studies ensued in marketing and information systems, only to present a complex picture with widely dispersed effect sizes and paradoxical findings even at the meta-analytic level: While some meta-analyses found strong evidence of the choice-overload effect and identified its moderators Chernev et al., 2015; McShane and Böckenholt, 2018), others contended a failure of replication and a general lack of empirical support for the effect (Scheibehenne et al., 2010; Simonsohn et al., 2014). These contradictory results leave the empirical understanding of choice overload in a fragmented state.

The fragmented views of choice overload directly result from the paradoxical combination of an obviously non-linear (i.e., inverted-U) conceptualization of the effect and a methodological deficit at the empirical front to examine such an effect, most prominently the lack of a meta-analytic method that can synthesize the results of two-group experiments to probe the characteristics of a non-linear relationship. Thus, via *Theore*, we premise the reconciliation of the fragmented views on developing a novel analytical link between the conceptual underpinning of choice overload -- i.e., the notion of the option-satisfaction relation being an inverted-U if choice overload exists, and monotonic if it does not -- with the unique analytical patterns discernible from the observed inconsistencies of the existing empirical findings. Ideally, such a link should explicate the mechanisms through which the former entails the latter. It would enable us to address the fragmented empirical view by elucidating how an underlying conceptual model could actually explain and account for the empirical inconsistencies. Similarly, we would reconcile the fragmented conceptual view by leaning on the coalescent of existing empirical evidence to statistically unpack the underlying option-satisfaction relation, determining whether it is monotonic or an inverted-U, and delineating the type(s) of moderation that should be expected.

More details about this case study are available via our working paper, which is under review at *Management Information Systems Quarterly*:

Zhang, N. and Xu, H. (2020). "Reconciling the Paradoxical Findings of Choice Overload Through an Analytical Lens," *under 2nd round review at Management Information Systems Quarterly*.

This case study demonstrates *Theore*'s capability of detecting heterogeneity caused by curvilinear relationships. The additional intricacies of curvilinear relationships (compared with the linear ones) give

rise to significant challenges in meta-analysis, where the notion of linearity underpinned the development of most existing methods. Indeed, there are not even adequate procedure to test for the presence of a curvilinear relationship in a meta-analysis of experimental studies. *Theore*'s novel analytical method of inferring the existence and moderation of an inverted-U relationship from two-group experiments is pertinent for at least three reasons:

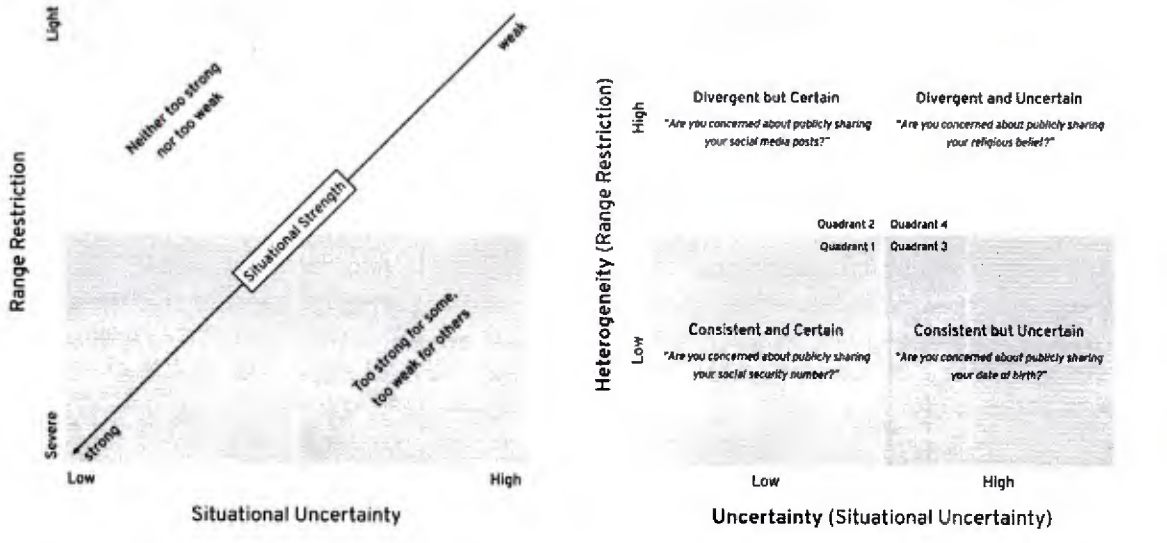
First, it allows researchers to tap the gold mine of existing empirical evidence in theorizing and testing an inverted-U relationship, even when the existing research designs were all based on the linear mechanism. Second, *Theore* enables the statistical disentanglement of the two types of inverted-U moderation, even when such moderation was never theorized in the literature nor brought to bear in the research designs. Last but not least, the findings of our method provide valuable guidelines for the design of future experimental studies. Specifically, when designing an experiment dedicated to the examination of an inverted-U relation, researchers may select a small set of "critical" independent-variable (i.e., X) values according to the estimations produced by *Theore*, instead of having to examine the full spectrum of X. For example, if *Theore* suggests a moderating effect that steepens/flattens the curve, researchers may want to select one value of X adjacent to the turning point and two values far away from the turning point to its left and right, respectively, in order to exemplify the change of steepness in the observed effect. On the other hand, if *Theore* suggests a moderating effect that shifts the curve left/right, researchers may want to select multiple values of X around the potential range of the turning point, in order to pinpoint the degree of shift conditioned by the moderation.

3.3. Case Study: Using *Theore* to Detect Heterogeneity Caused by Contextual Effects

In this case study, we aim to demonstrate *Theore*'s capability of detecting heterogeneity caused by contextual effects in privacy research. Many research studies have shown that people's privacy attitudes and concerns are often diverse, dynamic, and situation specific. Further, people's stated concerns about privacy, especially when elicited in a broad, general manner, often appear inflated when compared against their actual behavior in a specific context. For example, a person might report a high level of concern in response to a general survey question like "How concerned are you with online privacy?" yet openly share her personal diary online. This challenge is cited as a reason prompting researchers to increasingly conceptualize privacy in a fully contextualized manner, which calls for the measurement of privacy concerns to be implicated in a particular context.

While there is little doubt that contextualization is very important for understanding a highly "fluid" construct like privacy concerns, the limitations of such a context-specific approach to privacy research should also be noted. First, most context-specific privacy studies focused primarily on one or a few contexts (e.g., online shopping in one study, social media in another). This defers the comparisons between findings from different contexts, and the development of broad-range context-contingent theories, to the time when a sufficient number of context-specific studies have been accumulated to allow for theory-grounded meta-analyses to test such context-contingent theories. Second, in any context-specific study, researchers have to consider numerous contextual factors, like the type of private information involved (e.g., list of Facebook friends or online purchase records), the entity posing privacy threats (e.g., online advertisers or e-commerce companies), etc. Decisions as of how to contextualize a study, and along which factors, should be grounded in theory. Yet, in the extant literature, the identification of contextual factors is often done in a post hoc descriptive fashion (based on the data researchers have in hand) to fulfill journalistic imperatives, making it rather slow, if at all possible, to develop a theoretical framework guiding the selection of contextual contingencies in future research.

The objective of this case study is to examine the multiplicity of contexts and their impact on consumers' self-reported privacy concerns, identify challenges researchers may face in contextualizing privacy research, and offer possible ways forward. As shown in Figure 4, *Theore* helped us explicate how two pronounced effects of context, range restriction and situational uncertainty, alter the grounds on which people ascribe meanings to "privacy concerns". Through *Theore*, we quantitatively assessed the magnitude of heterogeneity and uncertainty occasioned by a given context.



(a) A Quadrant View of the Situational Strength of a Context (b) Conceptual Illustration of Our Two-Dimensional Framework

Figure 4. Using *Theore* to Detect Heterogeneity Caused by Contextual Effects in Privacy Research

This case study demonstrated *Theore*'s important contributions to the privacy literature. First, *Theore* quantitatively demonstrated how contextual factors can substantially shift the distributional properties of people's stated privacy concerns. The surprisingly wide dispersion of context effects substantiated the need for researchers to pay special attention to how they contextualize a privacy study, especially when examining the generalizability of their research findings across contexts. Second, *Theore* provided a diagnostic device for comparing and contrasting the effects of different contexts, so as to unpack the distributional properties of the self-reported privacy concerns, and impute from them two important dimensions of context-contingent shifts. It introduces an emerging method for survey data analysis with an easy-to-use visualization tool that provides important hints as to how context shapes people's self-reported privacy concerns and how it governs the conditioning of privacy concern-behavior relationships.

We recommend researchers consider applying *Theore* in pilot studies so as to identify potential caveats stemming from the context effects. Second, we recommend researchers demarcate the boundary conditions of their findings based on the context effects quantified by *Theore*. As shown earlier in the case study, whether to situate the elicited privacy concerns in the context of social media or e-commerce could considerably alter the meanings a respondent ascribes to "privacy concerns". Similarly, contexts may exert varying degrees of influences on the full range of attitudes and behaviors that are involved in a study. As a result, while a finding from a social-media context might readily generalize to the context of online ads, it might not apply as well to a context of e-commerce websites. *Theore* could help

researchers with demarcating the contextual boundaries of a study based on the quantified effects of the study context and the target one. More details about this case study are available via our working paper, which is under review at *Management Science*:

Zhang, N. and Xu, H. (2020). "From Contextualizing to Context-Theorizing: Assessing Context Effects in Privacy Research," *under review at Management Science*. Available at SSRN: https://ssrn.com/abstract=3624056

4. References

Beath, K. J. (2014). A finite mixture method for outlier detection and robustness in meta-analysis. Research synthesis methods, 5(4), 285-293.

Belkin, M., & Sinha, K. (2010, October). Polynomial learning of distribution families. In 2010 IEEE 51st Annual Symposium on Foundations of Computer Science (pp. 103-112). IEEE.

Berlin, J. A. (1995). Invited commentary: benefits of heterogeneity in meta-analysis of data from epidemiologic studies. American journal of epidemiology, 142(4), 383-387.

Böhning, D. (1999). Computer-assisted analysis of mixtures and applications: meta-analysis, disease mapping and others (Vol. 81). CRC press.

Breuer, C., Hüffmeier, J., & Hertel, G. (2016). Does trust matter more in virtual teams? A meta-analysis of trust and team effectiveness considering virtuality and documentation as moderators. Journal of Applied Psychology, 101(8), 1151.

Chernev, A., Böckenholt, U., & Goodman, J. (2015). Choice overload: A conceptual review and meta-analysis. Journal of Consumer Psychology, 25(2), 333-358.

Dasgupta, S. (1999, October). Learning mixtures of Gaussians. In 40th Annual Symposium on Foundations of Computer Science (Cat. No. 99CB37039) (pp. 634-644). IEEE.

De Jong, B. A., Dirks, K. T., & Gillespie, N. (2016). Trust and team performance: A meta-analysis of main effects, moderators, and covariates. Journal of Applied Psychology, 101(8), 1134.

De Guinea, A. O., Webster, J., & Staples, D. S. (2012). A meta-analysis of the consequences of virtualness on team functioning. Information & Management, 49(6), 301-308.

Fletcher, J. (2007). What is heterogeneity and is it important?. BMJ, 334(7584), 94-96.

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. The BMJ, 327(7414), 557-560.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index?. Psychological methods, 11(2), 193.

Hunter, J. E., & Schmidt, F. L. (2004). Methods of meta-analysis: Correcting error and bias in research findings. Sage Publications.

Hunter, J. E., & Schmidt, F. L. (2014). Methods of meta-analysis: Correcting Error and Bias in Research Findings (3nd ed.). Sage Publications.

Kalai, A. T., Moitra, A., & Valiant, G. (2016). Disentangling gaussians. Communications of the ACM, 55.

McLachlan, G. J., & Basford, K. E. (1988). Mixture models: Inference and applications to clustering (Vol. 84). New York: M. Dekker.

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. Perspectives on Psychological Science, 11(5), 730-749.

McShane, B. B., and Böckenholt, U. 2018. "Multilevel multivariate meta-analysis with application to choice overload," Psychometrika (83:1), pp. 255–271.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological bulletin, 105(1), 156.

Moitra, A., & Valiant, G. (2010, October). Settling the polynomial learnability of mixtures of gaussians. In 2010 IEEE 51st Annual Symposium on Foundations of Computer Science (pp. 93-102). IEEE.

Naaktgeboren, C. A., van Enst, W. A., Ochodo, E. A., de Groot, J. A., Hooft, L., Leeflang, M. M., ... & Reitsma, J. B. (2014). Systematic overview finds variation in approaches to investigating and reporting on sources of heterogeneity in systematic reviews of diagnostic studies. Journal of clinical epidemiology, 67(11), 1200-1209.

Nord, C. L., Valton, V., Wood, J., & Roiser, J. P. (2017). Power-up: a reanalysis of 'power failure' in neuroscience using mixture modeling. Journal of Neuroscience, 37(34), 8051-8061.

Paterson, T. A., Harms, P. D., Steel, P., & Credé, M. (2016). An assessment of the magnitude of effect sizes: Evidence from 30 years of meta-analysis in management. Journal of Leadership & Organizational Studies, 23(1), 66-81.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society of London. A, 185, 71-110.

Richard, F. D., Bond Jr, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. Review of General Psychology, 7(4), 331-363.

Romieu, I., Berlin, J. A., & Colditz, G. (1990). Oral contraceptives and breast cancer review and meta-analysis. Cancer, 66(11), 2253-2263.

Scheibehenne, B., Greifeneder, R., & Todd, P. M. (2010). Can there ever be too many options? A meta-analytic review of choice overload. Journal of consumer research, 37(3), 409-425.

Schlattmann, P., Verba, M., Dewey, M., & Walther, M. (2015). Mixture models in diagnostic meta-analyses—clustering summary receiver operating characteristic curves accounted for heterogeneity and correlation. Journal of clinical epidemiology, 68(1), 61-72.

Seim, K., & Sinkinson, M. (2016). Mixed pricing in online marketplaces. Quantitative Marketing and Economics, 14(2), 129-155.

Simonsohn, U. (2018). Two lines: a valid alternative to the invalid testing of U-shaped relationships with quadratic regressions. Advances in Methods and Practices in Psychological Science, 1(4), 538-555.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). p-curve and effect size: Correcting for publication bias using only significant results. Perspectives on Psychological Science, 9(6), 666-681.