Peer Review File

Quantifying Institutional Gender Inequality in Contemporary Visual Art

Corresponding Author: Professor Albert Barabasi

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
This study examines evidence of gender representational issues in the arts. The authors focus on two measures: gender-balance and gender-neutrality. Gender-balance is to expect institutions to have 50% of their exhibitions be women, while gender-neutral expects institutions to randomly draw exhibitors using the population of artists' gender (approximately 36.5% women). The authors analyze several institutions and find which follow gender-balance vs. gender-neutral strategies. The authors also analyze other inequalities in the art world, including exhibitions, auctions, and sales.

The article critically analyses this understudied topic using an extensive sample. Also, it offers dimensions of the disparities, going from opportunities to exhibit to sales. The authors highlight limitations in their estimation of gender, such as not covering non-binary gender identities, and their dataset coverage, such as not covering perhaps smaller venues with different gender representations. The authors also distinguish between population disparity and exhibition disparity. I have three topics that I would like the authors to discuss.

- What should galleries do? More value calling or discussion, or both, is needed. The authors propose two ways of solving the gender gap: gender-balance and gender-neutrality. But in the case of gender representation across artists (~36% women), you cannot meet both standards, at least statistically speaking. Are the authors willing to do a little bit more digging on a value call regarding this issue? The same goes for the analyses between museums and galleries.

- Changes over time? The author possesses a substantial sample over a long time. At least most of the data lies from 20 years ago. Do the authors see any sign of "improvement" in the art world?

Reviewer #2

(Remarks to the Author)
I am grateful for the chance to have reviewed this paper. It is extremely valuable and well-written, and I hope it is eventually published. I do have several points of concern and other comments that I will communicate below. Nonetheless, I commend the authors for taking a comprehensive approach to a big and thorny problem.

I will provide comments under the headings suggested in the Nature Communications reviewer guidelines (https://www.nature.com/ncomms/for-reviewers/writing-your-report), modified slightly. I will use consecutive numbering of comments that I believe should be addressed.

Key Results

This work takes a comprehensive look at the role of gender in contemporary art. The authors lay out a helpful framework for the study, namely the hypotheses of gender balance vs. gender neutrality. The authors study both exhibitions and auctions. Results derive from testing various measured gender ratios under these hypotheses, studying geographic sub-units of the data, examining co-exhibition networks, investigating the role of institutional prestige, looking at auction prices, and more.

Table 2 presents some primary measures extracted from the data. Some fascinating key conclusions from the paper include: (1) 54% of institutions are gender neural but only 30% are gender-balanced, (2) the fraction of man-preferred institutions is positively correlated with institutional prestige, and (3) co-exhibition gender is associated with a higher impact of success than the artist's own gender.

Significance

This work is highly original and very important. I know of no other study about gender diversity in the art world that provides empirical answers to so many key questions. I think this will be a highly influential paper, especially if it finds it way out of the usual science / social sciences readership and into the hands of curators, gallery owners, and other stakeholders.

Validity of Data and Analysis
----

(1) When I read the main manuscript, I wrote various comments about the need to adjust p-values. Then, I saw your comment in the SI that you chose to use unadjusted p-values. I am open to being convinced, but right now, I feel very concerned about this, perhaps because I did't really understand your explanation for that choice. Could you please provide more explanation and clarity so that I can understand the reasoning for your choice?

(2) I am also concerned about calculations done for non-US countries. In your assessments of gender, did you account for the extreme population gender imbalance in some of those countries? As merely one example, I see the United Arab Emirates in your study. This country is nearly 70% men.

(3) Another concern relates to your prestige classifications. You state very clearly the cutoffs you use (40th percentile and 70th percentile) to form three levels of prestige, but I would like to see justification for these. In other words, what does the distribution of prestige scores look like? If the prestige scores cluster into three groups, those could be a more natural way to label the institutions. If the distribution is very continuous, I worry that your results might be sensitive to those choices of 40th and 70th percentile, because you are essentially imposing false separation on data that is in reality close together. If you go the route of percentile cutoffs, there should be some sensitivity analysis. In other words, if one changes those cutoffs, do your basic results about the role of prestige stay the same?

(4) Similarly, you mention that you construct co-exhibition network for each artist with more than 10 exhibitions, but why 10?

(5) In your logistic regressions of Table 6, I think standard errors should be reported. More substantively, the paragraph at the bottom of page 17 causes me some confusion. On one hand you say you're not interested in predictive ability. On the other hand, you say that exhibitions per year and career length are dominant drivers, and I'd say that it's really hard to assess that if you haven't examined the predictive power of your model. At best, I think you can say that there are positive associations between those two factors and the probability of transitioning. If you want to say they are primary drivers, I think you would need to provide additional evidence. Also, because you fit (I believe) 21 parameters, did you adjust p values for multiple estimates?

(6) In the SI, you mention your use of genderize.io. I appreciate the mention of some of the limitation of the algorithm (including around Asian names). Still, I think you should provide more information and possibly some robustness checks. As we both know, genderize provides a probability of the inferred gender, and that probability is actually a frequency calculated as proportion of total occurrences within the underlying database of names. Your choice of a 0.6 probability cutoff feels unjustified to me, and I worry also about the admission of inferences associated with low occurrence counts (you don't mention using a cutoff for these). On one hand, I'd be comfortable with an unjustified cutoff for p if you used a much higher value, but on the other hand, doing so might exclude a lot more data, and possibly not in a gender-unbiased way. In any case, I think it is worth exploring the role that the choice of cutoff plays or somehow providing more justification.

Clarity and Context
----

(7) In the abstract, I think the reference to "machine learning" is overblown and/or misleading to most readers. In other words, unless I missed something, you have done logistic regressions. Please don't misinterpret my comment: I think that logistic regressions are a great and appropriate thing to do here. But they certainly existed well before the terminology of machine learning was around. Moreover, I think that in machine learning, the goal of logistic regression would be classification, and you explicitly say you are NOT trying to build a predictive model. In short, I would remove or at least drastically rephrase this.

(8) General comment for introduction: The audience of this journal might benefit from an explanation of the difference between a museum's permanent collection and its exhibitions. There are a lot of basic things about how art museums work that a general science and/or social sciences audiences will not know.

(9) On page 1, I might remove the word "quality" in your phrase about "nature of the quality of the art." For a gender disparity to be rooted in quality of the art requires that quality of the art is a thing that objectively exists, which I am not certain I

believe.

(10) On page 3 and throughout, I applaud you for addressing the limitations of using a binary coding of gender. Still, I would recommend going a step further and explicitly addressing/admitting that this may result in misgendering of individuals and/or erasing identities. To be clear, there's no way to do this research WITHOUT harm/erasure, and I am not suggesting that the research not be done. You could use the limitations of the data as a call for more inclusive gender data. You may have said this once but I think saying it right from the start of the paper is both appropriate (which is most important) and might protect you from visceral pushback (which is a secondary benefit).

(11) On page 4, when I got to this section, I was genuinely confused because I did yet not know what was to come later. As I read the main two paragraphs, I kept jotting notes like "this strikes me as the wrong comparison," and "why no statistical test?" At the end of the section you state that the goal is to move beyond these simple measures, but I wonder if your readers would be better served by putting that up front: "we're going to do something ludicrously simple and then we're going to move on to something more sophisticated."

(12) On page 8, you begin talking about prestige, but you only define the way prestige is calculated later on. I think this might be confusing for many readers and I would recommend rearranging or at the very least, foreshadowing.

(13) Also on page 8, you use the abbreviations "MoMA" and "MUMOK." Most readers will not know these, so you should define them. This also holds for the caption of Figure 2, where you mention MoMA.

(14) On page 9, in Figure 2, the horizontal axis labels in panel (e) run together and should be fixed.

(15) In the SI, on page 12, writing "0.00" for p values is not helpful. Can you please give more of a sense of the order of magnitude, or a coding with asterisks for different levels of significance, or anything else?

References

The references are appropriate.

Reviewer #3

(Remarks to the Author)
Summary: The authors consider the gender inequality in contemporary visual arts. In particular, authors distinguish two forms: gender balance (population proportions) and gender neutrality (Difference between the visibilities in exhibitions of those already in the populations). These metrics are also dissected based on institutional prestige to depict the intersectionality of institutional prestige and gender inequality.

Although gender inequalities have been explored in various fields and communities, this is the first effort that I am aware of that extended and effort to do so for the field of visual arts. Additionally, the datasets that the authors have made available will hopefully lead to more efforts continue this area of work. In that sense, I appreciate the authors' effort. However, I have several comments about this work which I elaborate below:

Comments:
1. Several previous works focussed on gender inequities in citations using very similar notions that the authors use (gender balance and gender neutrality). They also have have reached conclusions that agree with and strengthens the conclusions that the authors reach. In particular, the following three papers (in chronological order) are heavily related to the authors' work in my opinion:

(i) Dworkin, Jordan D., Kristin A. Linn, Erin G. Teich, Perry Zurn, Russell T. Shinohara, and Danielle S. Bassett. "The extent and drivers of gender imbalance in neuroscience reference lists." Nature neuroscience 23, no. 8 (2020): 918-926.

(ii) Nettasinghe, Buddhika, Nazanin Alipourfard, Vikram Krishnamurthy, and Kristina Lerman. "Emergence of structural inequalities in scientific citation networks." arXiv preprint arXiv:2103.10944 (2021).

(ii) Teich, Erin G., Jason Z. Kim, Christopher W. Lynn, Samantha C. Simon, Andrei A. Klishin, Karol P. Szymula, Pragya Srivastava et al. "Citation inequity and gendered citation practices in contemporary physics." Nature Physics 18, no. 10 (2022): 1161-1170.

(i) mentions "…..we find that reference lists tend to include more papers with men as first and last author than would be expected if gender were unrelated to referencing."
(ii) mentions "…female authors, who represent a minority of researchers, receive less recognition for their work (through citations) relative to male authors; second, authors affiliated with top-ranked institutions, who are also a minority, receive substantially more recognition compared to other authors…..we find that merely increasing the minority group size does little to narrow the disparities…."
(iIi) mentions "citation imbalance in favor of man-authored papers is highest for papers authored by men, papers published in

general physics journals and papers for which citing authors probably have less domain or author familiarity.”

The above findings (in particular from ii and iii) agrees with what authors find: merely getting rid of gender balance may not be sufficient to achieve gender neutrality (they have used definitions which normalizes for one factor while considering the other). Additionally, they also suggest strategies to mitigate the gender inequality via strategies that compliment the authors’ strategies. As such, I believe a discussion of such highly related work (though not on the field of visual arts) and how they align/differ from the authors’ conclusions will be a good addition to the paper.

2. The authors use the term “systemic gender inequality”. However, most of the authors’ analysis is descriptive in nature (via comparison of proportions, p-values, etc.) and the model they consider is probabilistic rather than causal. Could the authors shed more details as to how they justify the observed disparities being “systemic”? For example, the values in Fig. 1(a) for population and exhibition categories agree very closely with each other leading to the question “is there a causal relation between gender balance and gender neutrality and gender disparity in auctions?” If so, in which direction? I do not have an answer to offer here. But discussing these aspects (or at least highlighting the difficulties to explore them) will justify the observed disparity being “systemic”.

3. The authors have considered the network of institutes. Is there an analogous form of a network of artists which will reveal another aspect of the picture? E.g., in terms of the structure, do these two networks differ? If so, do those differences shed more light on the? This question occurred to me as a recent work which I came across (Lerman, Kristina, Yulin Yu, Fred Morstatter, and Jay Pujara. "Gendered citation patterns among the scientific elite." Proceedings of the National Academy of Sciences 119, no. 40 (2022): e2206070119.) used the difference between ego networks of authors to classify their gender, the point being that there is such a larger difference between networks of men and women so that an ML algorithm can use them to identify the gender. Maybe this could be helpful (at least for future work).

Reviewer #4

(Remarks to the Author)
This manuscript aims to understand the gender disparities in the world of contemporary visual arts. They found systematic gender differences in terms of artists, exhibition opportunities, auction opportunities, and auction prices. I found this study interesting and does provide first-hand evidence of the gender disparity issues in visual arts.

One concern for the analysis was that the artist productivity factor was not mentioned anywhere in the study, which could possibly interfere with the number of exhibits and auctions.

Page 2: I also found this argument problematic. The manuscript states, “Studies show that people cannot infer the gender of the artist from the work itself [5, 2], suggesting that the observed gender disparity cannot be rooted in the nature or the quality of the art men and women produce. As Nochlin hypothesized, the answer is likely rooted in ‘the very nature of our institutional structures.’”. My understanding here is that if, as claimed by the manuscript, people are unable to infer the gender of the artists, then why would it suggest “that the observed gender disparity cannot be rooted in the nature or the quality of the art men and women produce”? I had a hard time making the connection.

Figure 4: I found it hard to reach the conclusions the manuscript claimed based on Figure 4a based on direct observation. (1) What color was used for the edge connecting a man-preferred and woman-preferred institution? The choice would have an impact on the visual impression of the network. (2) Nodes in the more peripheral locations do seem to show inequality-based assortativity, but it is hard to tell about the large proportion of the nodes in the center of the network. I think the reader would benefit from additional indicators or different network representations to support the claim.

Figure 5 needs legends.

Reviewer #5

(Remarks to the Author)
Comments on:

Quantifying Systemic Gender Inequality in Contemporary Visual Art

Overview

While the most interesting scientific papers clearly identify a hypothesis to be tested, typically derived from a particular theoretical framework, and proceed to test the hypothesis using observed data, there are potential contributions to be made also by descriptive papers that measure a phenomenon or relationship with care and present new and unexpected findings that can help encourage development of theory or identification of hypotheses to be tested. This paper could potentially be put into this second category. Providing, as suggested by the title, a quantitative measurement of gender inequality in contemporary visual art. Presumably this would mean the art market, consisting of the production, consumption, sale and final prices of art works.

The central findings reported in this paper are:

1. Combining both museums and commercial galleries into the category of "art institutions" almost 54% are gender neutral (exhibiting artists in proportion to the gender share in the population of contemporary artists). About 30% are gender balanced (half of artists exhibited from each gender).
2. Among those that are not gender-neutral, museums are split almost evenly between favoring women versus favoring men, while commercial galleries split almost 2 to 1 favoring men over women artists.
3. Using a notion of "prestige" for exhibition venues determined in an earlier paper, higher prestige institutions tend to favor male artists. This difference is very small based on gender neutrality, but larger based on gender balance.
4. Of the 65,768 artists analyzed, 24% of the men and nearly 16% of the women had at least one auction sale. In terms of total value of auction sales, the value of works sold by men artists was over 7.5 times that of women artists although most of this difference is due to the number of objects sold rather than the average price of the objects. The average sales price of an object sold by men was 1.75 times the average price of a work by a woman artist.
5. A higher proportion of museums are gender-neutral than of commercial galleries (60% versus 50%), although it should be noted that over 60% of art institutions, accounting for 11% of exhibitions have too few exhibitions to be reliably categorized.

Somewhat less successfully, the paper puts forward some additional claims. I will explain below what I see as problems with these results, but I list them here because the authors put them forward as central findings.

6. Art institutions with comparable degrees of gender inequality tend to be connected in a directed exhibition network to each other, often exhibiting the same artists.
7. Artists (whether male or female) may have careers that result in their exhibitions being at predominantly man-preferred, woman-preferred, or neutral exhibition venues. Comparing an artist's first five exhibitions to their last five exhibitions reveals considerable stability to this exhibition proclivity, and the tendency to stay in the same category is somewhat stronger for artists who exhibit at low-prestige venues.

Novelty of results and review of related scholarship

The problem of gender inequality in the art world is an issue that has received considerable attention from other scholars whose results have been published previously. Some of these are cited as old versions and should more properly be cited as the versions that have passed peer-review. This includes Adams, et al. (2021), Bocart, et al. (2021) and Cameron, et al. (2019) (full citations given below). Other relevant and important contributions are not cited at all, for example Cowen (1996), Droege (2022), Galenson (2007), Greenwald (2021), Hoffman & Coate (2020), and LeBlanc & Sheppard (2022).

There are occasions in the paper when citations are offered that seem to betray a misunderstanding of the contents of the cited work, or where the cited work seems inappropriate. For example, at the bottom of page 4 the manuscript asserts that pressure to address gender imbalance "… stems from the strong belief and evidence that artistic performance and artistic aptitude are agnostic to gender." A citation is given to Bourdieu and to the work of Adams, et al. (2021). Bourdieu's beliefs are often inchoate and vague, and certainly his 1993 collection of essays presents almost no evidence or formal analysis concerning gender and artistic skill1. Beliefs could be cited on both sides of the issue (one might as well include Georg Baselitz's "belief" that women don't paint very well).

Adams and her collaborators do address this issue, but come to a different conclusion from the one asserted. To be "agnostic to gender" would imply that works of art convey no information about the gender of the artist who created the work. Yet Table 4 and Figure 1 in Adams, et al. (2021) show that using titles as indicators of image content, they can differentiate between artworks produced by men and by women. In the final section of their paper, they do show that an online survey of small to modest samples representative of the general US population are mostly unable to reliably determine the artist gender based on online images. They note that the results won't generalize if this sample is not representative of art collectors or art auction participants, and their conclusion is that for this sample representative of the entire US population, "… structural differences that might exist are not readily observable."

Differences between artworks by men and by women are also noted in the paper by LeBlanc and Sheppard (2022). Their analysis shows that women artists produce works that are smaller (about 16% smaller in area) and less chromatically diverse images, that are less likely to be portraits or landscapes, or dated, or produced using oil paints on canvas. Interestingly, their analysis uses the Google SafeSearch image analysis neural network to show that women artists are more likely to produce images classified as likely to contain Adult or Racy images.

So while there is "belief" existing on both sides of the issue of whether artworks are "agnostic" to gender, the actual empirical evidence is at best mixed and in my view actually support the assertion that art works by women DO differ in some important ways from works by men. A more careful and complete reading of published research on the subject would improve the exposition of this paper.

As suggested by both the works cited in this paper and the additional works mentioned above, the problem of under-representation of women artists in exhibitions, auctions, and sales is widely recognized and has been the subject of scholarly inquiry for over 50 years. The findings of this paper that provide quantitative confirmation of this are therefore not really novel or surprising. If anything the gender gaps that are identified in this paper are somewhat smaller than might have been expected given the large volume of scholarship as well as journalism and commentary on the topic. This makes central findings 1 through 4 listed above useful confirmation, but not a significant contribution to the literature. This literature, indeed, has been "moving on" to more carefully test hypotheses about why these gender gaps exist. Thus for example Adams, et al. (2021) test the importance of local social attitudes towards women as an important determinant of the gender gap. LeBlanc and Sheppard (2022) consider the intersection of gender and ethnicity and different characteristics of the works themselves. Greenwald (2021) by contrast, looks at supply constraints arising from limited access of women artists to formal training or the demands of family and personal life as a factor.

The current manuscript has one result that seems as if it could provide an avenue for a test that has not been emphasized by other scholars: the difference in gender gap between museums and commercial galleries listed above as central finding 5. This is novel in part because very few scholarly analyses have access to data on outcomes in commercial galleries, relying almost exclusively on art auction sales instead. Galleries are businesses who, in pursuit of their objective to earn profits for their owners, must cater to the desires of collectors and others who are their customers. Museums, by contrast, are in almost all cases government or non-profit organizations with more complicated missions to preserve and care for objects, to attract and educate the public, while still generating sufficient revenues to “keep the doors open”. While not conclusive, finding 5 might be generally interpreted as supporting the idea that the gender gaps observed in the art market are the product of “demand side” preferences of collectors and buyers rather than supply constraints.

Unfortunately, the paper cannot or does not test this potential hypotheses, choosing instead to regard galleries and museums as essentially equivalent. Virtually every mention of one type of institution, in both the main article text and the supplemental material, is within one word of the other type separated only by the word “and”. While the general public may make this mistake, it really is inexcusable for scholars of the art market to conflate the two. While both of these institutions stage exhibitions and artists are anxious to be included, listing exhibitions of both types on their CVs or web sites, they are not equivalent. In general, artists would prefer to have shows at museums and would prefer for their work to be sold to museums. I have personally had the experience of working with museums seeking to acquire the work of an artist from a gallery and, once it is know that the sale would be for a museum collection an immediate discount in price of 10% is offered. To regard the two types of institutions as equivalent is a fundamental error in understanding the art market and misses a real opportunity to illuminate a possible source of the gender gap.

Network and data analysis methodology

This fundamental asymmetry between the types of institutions included in the sample raises further difficulties for the network analysis and measurement of “institutional prestige” that is offered as an important variable in understanding the exhibition gap between men and women artists. The first problem that arises is that with exhibition venues as the nodes and the strength of links or flows between nodes determined by the number of artists exhibiting first in one venue then in the other, the intrinsic difference in the institutions means that a link from a gallery to a museum should not be regarded as making the same contribution to institution prestige as a link from a museum to a gallery. The former (say an exhibit at Hauser & Wirth followed a few years later by a show at the Tate Modern) is a signal of a successful artist whose career is building. By contrast the latter type of link (say an exhibit at Marian Goodman or Paula Cooper a few years after the Tate exhibition) may not say much about the artist, but more about the gallery taking advantage of the museum exhibition to generate profitable sales to collectors seeking to acquire a blue chip artist’s work. These nuances are missed by the network methodology as it is applied.

A related difficulty arises when, as in the results presented in Figuires 2c and 2d and discussed on page 10, the authors try to use the “prestige” of the exhibition venues as an explanatory variable for the fraction of venues that are gender neutral or gender balanced. The network of art exhibition venues with links defined as the authors have done presents a classic example of a network with endogenous link strength. Such networks have been extensively researched in the economics literature and the authors might benefit from a careful study of papers such as Bloch & Dutta (2009) or Olaizola & Valenciano (2020).

Of course, if the link strength is endogenous then the eigenvalue centrality will also be endogenously determined by the choices and investments made by the individuals (artists or gallery owners) and organizations (museums) that comprise the network. I’m sure I don’t have to explain to the data scientists among the list of authors that this endogeneity creates potential statistical bias in estimating or displaying the relationship between the “prestige” as measured by network centrality and a potentially noisy outcome like the fraction of venues that are gender neutral. While the existing paper does not emphasize estimates of parameter values using prestige rank as an explanatory variable, the presentation and comparison of slopes in Figures 2c and 2d are essentially the same thing and subject to the same problems of bias due to endogeneity.

I also that it curious that the authors grouped exhibition venues together into “low”, “mid” and “high” prestige groupings with dividing lines at the 40th and 70th percentiles. First one might ask why group them at all? This is not done, for example, in the 2018 Science companion paper (see for example Figure 2B on page 827). Grouping them has the effect of suppressing the ability of the reader to perceive noise in the data that can help evaluate the strength of the purported relationship. The situation is not helped by choosing unorthodox percentiles to create the groups and not offering explanation. Why not create the groups based on terciles of the prestige ranking data?

Finally, I found the presentation of the Co-exhibition networks in Figure 4, along with the claim on page 11 that the apparent “… emergence of clusters dense in blue or red nodes is evidence of inequality-based assortativity, indicating that institutions with comparable degrees of inequality tend to be connected to each other, often exhibiting the same artists” to be potentially misleading and inaccurate. Simply looking at the Figure the reader might reasonably get the impression of a random assemblage of red and blue dots and lines. We are told on page 10 that the network is printed using a force-directed algorithm, but no other details are provided. Producing meaningful printouts of complex networks is a notoriously difficult problem, and while force-directed algorithms of one sort or another are generally used, there are several specifications for the attractive forces and repulsive forces for the nodes that can affect what clusters appear to the eye of the beholder. To assert that there is an “emergence of clusters” feels like the authors trying to use a “Jedi mind trick” to convince the reader that there is something present that they are unable to establish through ordinary data analysis.

I did find the exploration of the idea of defining the co-exhibition gender of artists, and the potential change of co-exhibition gender, potentially interesting. There is again the problem of endogeneity of the network that is used as the basis of the presentation in Figure 5d and the inference that there is a difference in co-exhibition transition probabilities depending on low, mid, or high prestige exhibition venues, but the idea is novel and potentially interesting. A main difficulty is that it is not clear whether the co-exhibition gender is related to an outcome that we really care about. It is not clearly related to actual gender of the artist (a male artists could spend his entire career as a co-exhibition female, or a woman artist her entire career as a co-exhibition male) and therefore not related in any simple way to the gender gap in exhibition frequency.

The analysis presented suggests that co-exhibition gender, or interaction between actual artist gender and co-exhibition gender, may play a role in affecting the probability that an artist ever has a work offered for sale at auction. Unfortunately the failure to include any other controls (such as medium, support, or image content) as factors in determining appearance at auction, or the failure to explore the relationship between these variables and actual auction price make it difficult to tell if we (or if artists) should really care about this variable. Still, I would encourage the authors to continue to explore ways in which this measure might be a useful metric for understanding the importance of gender in the art market.

Other annoyances

While I have listed my major concerns about the paper above, there are a few other issues that I would recommend the authors address or at least consider as having caused at least one reader some concerns about the scientific quality of the research.

• The analysis focuses on about 65 thousand artists out of over 465 thousand – and a fully convincing explanation for constructing the sample is not provided. Why, for example, is consideration limited to those artists who began exhibiting on and after 1990. The explanation offered implies that this was required for accurate reconstruction of exhibition history, but I find this difficult to fully accept. The approach introduces potential sample selection bias into the analysis that should be acknowledged and discussed. Does this limit exclude Damien Hirst, for example (whose work featured in a group exhibition in 1988)? It surely would exclude Jeff Koons and many other “blue chip” contemporary artists, so some further explanation seems warranted.

• The current text contains numerous problematic assertions about how exhibition venues respond to incentives to make their decisions. For example, on page 6 it is asserted that the quantitative analysis of gender representation could be offered to the gallery owners, museum curators and boards, then it would offer “… them the metrics and structural insights to empower them to offer balanced representation.” I find it very difficult to believe that these decision makers lack the power change gender composition of the groups of artists they exhibit, nor that these quantitative measures will provide “metrics and structural insights” to these persons. It would be much better to think through a careful model of their decision making, use the metrics to test, refine, and validate the model, and then use the model to identify how to induce changes in their decisions.

• On page 15 the text states that “Sales … and the resulting valuation of an artist’s work, is a frequently used (and misused) measure of artistic success.” This kind of “nudge and wink – do you get it?” argument appears frequently in journalism, but in my view has no role in scientific writing. “Misused” in what sense? Why is focus on the value of an artist’s work a “misused” measure? Are artists supposed to not be paid? Is the assertion that for an artist to be “starving” is the natural order of things, or that for some reason artists are not supposed to be concerned about the valuation of their work, but only about the inner glow that they experience in its creation? The text doesn’t say, but simply asserts that focus on valuation is frequently misused. If the authors have evidence of this, present it. If they have a clear argument to make or to test using their data, tell us.

• Another example of what might be characterized as “unsupported or arm-chair social theorizing” is presented in the penultimate paragraph of the paper, where the authors assert that the gender gap in exhibition of artworks is “can be only balanced out through the adoption of affirmative acquisition and perhaps gender affirmative programming – i.e. policies whereby institutions over-emphasize art by women artists in their exhibitions and purchases aiming to move them closer to gender balance over time.” Do the authors have any evidence to support this claim? Any examples where such policies have been successful? I would suggest avoiding ending the paper with a claim that feels “progressive and relevant” but is unsupported by the evidence assembled or analysis presented. Such writing may be appropriate for Artnet but is probably best left out of serious scientific journals.

• The current text states on page 23 that the data “… with artist gender and institutional measures of gender inequality [is] now available on the GitHub repository ….” (emphasis added). As far as I can tell the GitHub repository does not contain the data, but only a 2-line readme file with no additional information.

References

Adams, R. B., Kräussl, R. Navone, M., & Verwijmeren, P. (2021). Gendered prices. The Review of Financial Studies, 34(8), 3789–3839.

Bloch, Francis & Dutta, Bhaskar (2009). Communication Networks with Endogenous Link Strength

Games and Economic Behavior, 66:39-56.

Bocart, F. Y., Gertsberg, M., & Pownall, R. A. (2021). An empirical analysis of price differences for male and female artists in the global art market. Journal of Cultural Economics, 1–23.

Brown, T. W. (2019). Why is work by female artists still valued less than work by male artists? www. artsy.net/article/artsy-editorial-work-female-artists-valued-work-male-artists.

Cameron, L., Goetzmann, W. N., & Nozari, M. (2019). Art and gender: Market bias or selection bias? Journal of Cultural Economics, 43(2), 279–307.

Cowen, T. (1996). Why women succeed, and fail, in the arts. Journal of Cultural Economics, 20(2), 93–113.

Droege, J. (2022). First impression biases in the performing arts: Taste-based discrimination and the value of blind auditioning. Journal of Cultural Economics volume 46, pages 391–437.

Galenson, D. W. (2007). Who were the greatest women artists of the twentieth century? a quantitative investigation. NBER Working Paper, No. w12928.

Greenwald, D. S. (2021). Painting by numbers: Data-driven histories of nineteenth-century art. Princeton, NJ: Princeton University Press.

Hoffman, R., & Coate, B. (2020). Fame, what's your name? Quasi and statistical gender discrimination in an art valuation experiment. Working paper.

Huhn, Tom (1993). The Field of Cultural Production: Essays on Art and Literature. Journal of Aesthetics and Art Criticism January 1993.

LeBlanc, A. & Sheppard, S. (2022). Women artists: gender, ethnicity, origin and contemporary prices. Journal of Cultural Economics, 46:439–481.

Olaizola, Norma & Valenciano, Federico (2020). Characterization of Efficient Networks in a Generalized Connections Model with Endogenous Link Strength, SERIEs, ,11:341-67, September 2020.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Thank you for your detailed responses to my comments. I appreciate the additional analyses.

However, I believe there is an opportunity for further discussion about what galleries and museums might do to address the gender gap. While I understand that your aim is not to prescribe specific actions for these institutions, your study clearly highlights a worrying pattern that needs addressing.

This addition would enhance the practical relevance of your research and might serve to guide those in the art world. I would suggest you reconsider adding a section discussing possible approaches that institutions could adopt or explore further to address the gender gap you have highlighted in your study. To that end, I recommend considering the findings and recommendations of the following papers, which, while not directly related to the art world, offer valuable insights into successful strategies for reducing gender biases and promoting equality:

- Carnes, M., Devine, P. G., Manwell, L. B., Byars-Winston, A., Fine, E., Ford, C. E., ... & Sheridan, J. (2015). Effect of an intervention to break the gender bias habit for faculty at one institution: a cluster randomized, controlled trial. Academic Medicine: Journal of the Association of American Medical Colleges, 90(2), 221.
- Kossek, E. E., Su, R., & Wu, L. (2017). "Opting out" or "pushed out"? Integrating perspectives on women's career equality for gender inclusion and interventions. Journal of Management, 43(1), 228-254.
- Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., ... & Greenwald, A. G. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. Proceedings of the National Academy of Sciences

Reviewer #2

(Remarks to the Author)
I have taken a look at the revision and at the very detailed responses of the authors. I appreciate the careful attention to the feedback and I am convinced by the revisions. I'd be very happy to see this important paper published.

Reviewer #3

(Remarks to the Author)
The authors have adequately addressed my comments. I do not have any further major comments.

Reviewer #4

(Remarks to the Author)
I appreciate the authors' efforts to address some of the comments.

I have the following additional comments/concerns.

(1)With the additional references added to the revised manuscript, I am not convinced that the findings of this study are novel enough. There have been similar studies with similar findings for the art world. The major difference between this main study and others might be that this study used a larger dataset than many of them.
(2) For the data, I'm concerned about the selection of artists included in the analytical sample. The manuscript indicated that it identified the gender category of 471,791 artists, but only kept 86,894 (18%) of them after removing individuals who started their careers after age 50 or before age 18. The study was then further limited to those who started their exhibitions on and after 1990. It was not reported in the study regarding the accuracy of the demographic information. It could be possible that those excluded exhibit certain patterns by gender. The further exclusion of artists who started on or after 1990 might suffer from incomplete coverage of individual careers, and their pattern based on gender (if any).
(3) I'd suggest avoiding using words to imply causality in the manuscript. For instance, on Page 3: "This focus is motivated by our goal of identifying systemic influences of gender representation on active artist careers;". There are many other similar instances throughout the manuscript.
(4) For the disparities in exhibition opportunity, the manuscript indicates the gender ratio is 1.86, the larger the gender ratio in the whole sample (1.74). Yet the mean difference between the exhibition opportunities of men and women is 0.9 (6% difference). Because productivity is not accounted for here, it is impossible to estimate how much of these disparities are due to productivity. It is also not clear here how sensitive the conclusion is to certain individual artists being extreme cases (i.e., certain super popular artists with many exhibitions).
(5) Maybe the concept of gender-balance and gender-neutrality should be introduced early on.
(6) In the analyses based on institutions, the manuscript indicated that 62.14% of institutions were excluded due to the limited number of exhibitions (n<30), which is about 10% of the total. What's the gender composition of those exhibitions? If larger/more prestigious institutions are more likely to favor men, whether/how does the excluding of these exhibitions speak to the conclusions here?
(7) For the regression analyses on auction opportunities, I find the statement "Overall, we find that the gender preference of institutions embracing an artist's work is a stronger indicator of the likelihood that an artist will transition to the auction market than the gender of the artist, capturing an institutionally driven gender premium for men" a bit disturbing. I don't think coefficients of different independent variables/models can be directly compared.
(8)The references seem to be messed up: the first reference is numbered 29. I thought the references appear in number order?

Version 2:

Reviewer comments:

Reviewer #4

(Remarks to the Author)
I appreciate the detailed responses and additional analysis in the revision. I don't have further comments.

Version 5:

Reviewer comments:

Reviewer #6

(Remarks to the Author)
I was invited to focus on the Bayes factor and sensitivity analysis part of the manuscript. After reviewing the manuscript and the associated GitHub scripts regarding the Bayesian Factor (BF) analyses, I have two concerns requiring clarification:

1. Prior Justification
The authors selected a beta(1,1) prior for their analysis but did not justify this choice. Given the critical role of priors in Bayesian inference, a rationale for selecting this uniform prior—or a sensitivity analysis comparing alternative priors (e.g., beta(0.5,0.5) or weakly informative priors)—should be included to support the robustness of the conclusions.

2. Classification Sensitivity and Simulation Transparency
The authors categorized institutions using a BF10 threshold of 1/3 and stated in their response that "with five exhibition observations, institutions are correctly classified as gender-neutral 85% of the time." However, there were two issues:
2.1 Documentation Gaps: The GitHub code does not explicitly include the simulation script to generate this 85% accuracy

claim. To validate this, I attempted to replicate a simplified scenario:
• Ground truth: 50% male artists (p = 0.5), n = 5 exhibits.
• Simulation result: 0 of 1000 trials met the BF10 < 1/3 threshold for correct classification (true negative); and 4 of 1000 trials met the BF10 >= 3 threshold (false positive)

2.2 Effect Size and "hit rate": The above simulation used the parameters from the model of H0. It is also important to simulate the scenario with H1. I tried two effect sizes (p = 0.8 and p = 0.55), with n = 5, the percentage of trials that met the BF10 > 3 threshold for the correct classification was 31% and 6.2%, respectively. Overall, my simulation suggests n = 5 is too small to detect either the true negative or true positive events.

To ensure reproducibility and validate the claimed performance, the authors should:
• Share the complete simulation code used for accuracy calculations.
• Report sensitivity analyses across varying n (number of exhibits) and effect sizes (i.e., the deviation of p from 0.5 or 0.364).

Below is my simulation code for reference:

```
# the code was based on JASP team's R code, the results are similar to the authors' (slightly different precision)
def bayes_factor(counts, n, a=1, b=1, theta0=0.5):
logBF10 = betaln(counts + a, n - counts + b) - betaln(a, b) - counts * np.log(theta0) - (n - counts) * np.log(1 - theta0)
BF10 = np.exp(logBF10)
return BF10

n = 5
p = 0.5 # also simulated with 0.365, 0.55, 0.8
x = np.random.binomial(n, p, 1000)

y = bayes_factor(x, n)
proportion = np.sum(y < 1/3) / 1000
print("Proportion of Bayes factors < 1/3:", proportion)
```

**
After attempting to reproduce the full code and re-read the methods and results sections, I have two comments for the frequentists statistics in the results section:

(1) On page 8, the comparison between museums and galleries is only presented as descriptive data; no statistical test was presented here. I am not sure the authors can claim one is higher than the other, e.g., "3.76% of galleries vs 2.55% of museums" does not necessarily demonstrate a "higher fraction of galleries ... than museums".

(2) In the series of logistic regressions, the model comparison index used was BIC, but no numeric results were presented showing the magnitude of BIC reduction.

As the data is available now, I was able to run most of the script, but I tried for a while to get the files in the right folder and create new folders with the specified names, yet I could not run all the code. Here are some suggestions regarding the provided script and data: first, I recommend the authors share the data along with the code, which will make it easier for future readers to reproduce the analysis—I personally do not believe there is a huge difference between sharing via DataVerse and GitHub; second, I recommend the authors add info regarding the computational environment used for the analysis (i.e., version of Python and Python modules, operating system); third, more comments are needed to navigate readers, especially explaining the different pipelines; finally, while some prefer Python scripts for speed, Jupyter Notebooks are more reader-friendly for understanding the data analysis flow when the data volume isn't extremely large.

In general, I recommend not using two systems of hypothesis testing to avoid contradiction (see references below). In this manuscript, the Bayesian and frequentist statistics are used for testing different hypotheses, which I find acceptable. For the part where statistical testing was missing (‘3.76% of galleries vs 2.55% of museums’), I also recommend using BF. If the authors are interested in using Bayesian analysis for the series of logistic regressions, they may consider using Bambi, a powerful Python module for Bayesian statistics.

Dienes, Z. (2024). Use one system for all results to avoid contradiction: Advice for using significance tests, equivalence tests, and Bayes factors. Journal of Experimental Psychology: Human Perception and Performance, 50(5), 531–534. https://doi.org/10.1037/xhp0001202
Schreiner, M. R., & Kunde, W. (2025). A cautionary note against selective applications of the Bayes factor. Journal of Experimental Psychology: General, 154(1), 294–304. https://doi.org/10.1037/xge0001666

(Remarks on code availability)
I reviewed the code for the beta-binomial model for BF (https://github.com/Barabasi-Lab/artGender/blob/main/codes/03-03-gender_preference_scatter_boundary.py), it seems to me that only partial script was shared, because the results of BF were

loaded from json files ("gender_neutral_country_bf10") that were not shared on github.

Version 6:

Reviewer comments:

Reviewer #6

(Remarks to the Author)
The authors have carefully addressed all my concerns, I have no further comments.

(Remarks on code availability)

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This study examines evidence of gender representational issues in the arts. The authors focus on two measures: gender-balance and gender-neutrality. Gender-balance is to expect institutions to have 50% of their exhibitions be women, while gender-neutral expects institutions to randomly draw exhibitors using the population of artists' gender (approximately 36.5% women). The authors analyze several institutions and find which follow gender-balance vs. gender-neutral strategies. The authors also analyze other inequalities in the art world, including exhibitions, auctions, and sales.

The article critically analyses this understudied topic using an extensive sample. Also, it offers dimensions of the disparities, going from opportunities to exhibit to sales. The authors highlight limitations in their estimation of gender, such as not covering non-binary gender identities, and their dataset coverage, such as not covering perhaps smaller venues with different gender representations. The authors also distinguish between population disparity and exhibition disparity. I have three topics that I would like the authors to discuss.

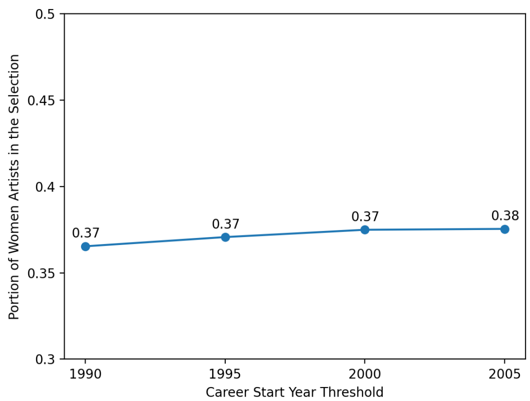
We thank the Reviewer for acknowledging the novelty of our study and of the extensive data empowering our analysis.

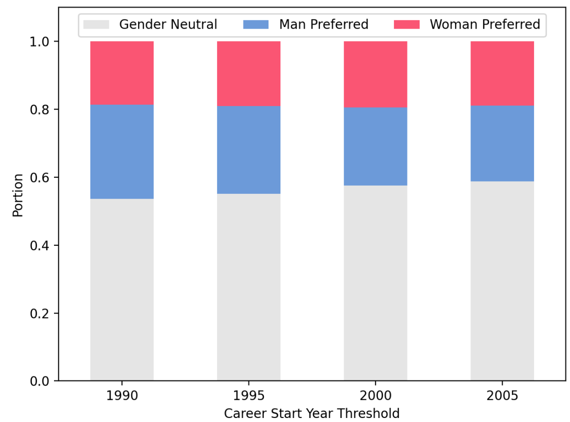
- What should galleries do? More value calling or discussion, or both, is needed. The authors propose two ways of solving the gender gap: gender-balance and gender-neutrality. But in the case of gender representation across artists (~36% women), you cannot meet both standards, at least statistically speaking. Are the authors willing to do a little bit more digging on a value call regarding this issue? The same goes for the analyses between museums and galleries.

We want to clarify that the proposal of the gender-balance and gender-neutral criteria are not methods to solve the gender gap, but are two perspectives for defining what constitutes a gender gap. As we discuss in the manuscript, ideally the art world should achieve gender-balance, but that necessarily means that it must first achieve gender-neutrality on route. It is beyond the scope of our work to suggest specific actions for galleries and museums to mitigate gender inequality, however we believe that our analysis of artistic careers and the interdependencies between them, makes a significant contribution towards understanding gender bias in the art world. As we have repeatedly seen in corporate and scientific environments, the first step towards designing appropriate policies and interventions to establish equity is developing a robust quantitative understanding of the extent of gender inequality.

- Changes over time? The author possesses a substantial sample over a long time. At least most of the data lies from 20 years ago. Do the authors see any sign of "improvement" in the art world?

We thank the Reviewer for focusing our attention on this important question. Following their recommendation, we conducted a sensitivity analysis with respect to the data processing decision of using 1990 as the cut-off year. Specifically, we increased the threshold for career selection from 1990 until 2005. As shown below (and also in the SI, Figure S2), both the proportion of women artists and the institutional classification under the gender-neutral hypothesis stay nearly constant for all thresholds, with a slight trade-off of some man-preferred institutions being reclassified as neutral.





In summary, we thank the Reviewer for their insightful comments which motivated us to revise our manuscript to identify gender-balance and gender-neutral as two perspectives for defining what constitutes a gender gap and to perform additional analysis of temporal changes in gender biases in art.

Reviewer #2 (Remarks to the Author):

I am grateful for the chance to have reviewed this paper. It is extremely valuable and well-written, and I hope it is eventually published. I do have several points of concern and other comments that I will communicate below. Nonetheless, I commend the authors for taking a comprehensive approach to a big and thorny problem.

We thank the Reviewer for their strong endorsement and for recognizing the value of our contribution, and greatly appreciate the detailed comments.

I will provide comments under the headings suggested in the Nature Communications reviewer guidelines (https://www.nature.com/ncomms/for-reviewers/writing-your-report), modified slightly. I will use consecutive numbering of comments that I believe should be addressed.

Key Results

This work takes a comprehensive look at the role of gender in contemporary art. The authors lay out a helpful framework for the study, namely the hypotheses of gender balance vs. gender neutrality. The authors study both exhibitions and auctions. Results derive from testing various measured gender ratios under these hypotheses, studying geographic sub-units of the data, examining co-exhibition networks, investigating the role of institutional prestige, looking at auction prices, and more. Table 2 presents some primary measures extracted from the data. Some fascinating key conclusions from the paper include: (1) 54% of institutions are gender neural but only 30% are gender-balanced, (2) the fraction of man-preferred institutions is positively correlated with institutional prestige, and (3) co-exhibition gender is associated with a higher impact of success than the artist's own gender.

Significance

This work is highly original and very important. I know of no other study about gender diversity in the art world that provides empirical answers to so many key questions. I think this will be a highly influential paper, especially if it finds it way out of the usual science / social sciences readership and into the hands of curators, gallery owners, and other stakeholders.

We are delighted by the reviewer's endorsement of our work and assessment of its potential importance to a wide variety of stakeholders.

Validity of Data and Analysis
----

(1) When I read the main manuscript, I wrote various comments about the need to adjust p-values. Then, I saw your comment in the SI that you chose to use unadjusted p-values. I am open to being convinced, but right now, I feel very concerned about this, perhaps because I didn't really understand your explanation for that choice. Could you please provide more explanation and clarity so that I can understand the reasoning for your choice?

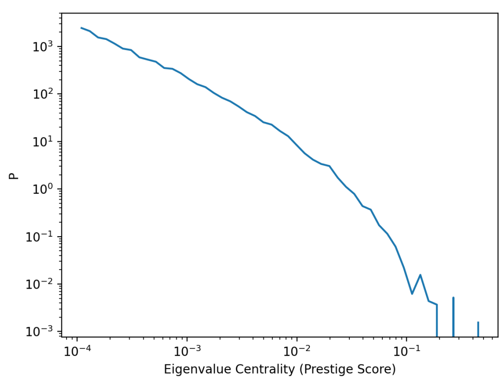
We thank the Reviewer for their careful consideration of the statistical significance of our institutional classifications. To reiterate, our ultimate goal is to elucidate how the interconnected structure of the art world perpetuates gender inequalities, and to quantify the impact on artists' careers. The starting point of this analysis was to classify art institutions based on the gender composition of their gender histories. The simplest way to do that would have been to threshold the fraction of exhibitions by women artists, such that we classify institutions above/below the threshold as woman-preferred / neutral / men-prefered. However, this threshold based approach misses the influences of the number of exhibitions (e.g. the number of samples). Therefore, we recast the problem of classification as a statistical hypothesis test to establish a principled way to vary the threshold based on the number of exhibitions. Note that these classifications are insightful regardless of whether they happened by chance or through the curator's explicit decision making criteria, the potential impact on artists in limiting access to future resources and rewards remains the same. Secondly, as the Reviewer mentions, in SI Section 3A we performed the analysis employing the Bonferroni correction, an extremely conservative method for adjusting p-values which typically results in an increase of type-II errors. We found that qualitatively the Bonferroni correction shows very similar results to the unadjusted results we report in the main manuscript. Therefore, we are confident that the organizing principles we uncover are not the product of random fluctuations in the gender preference classification.

(2) I am also concerned about calculations done for non-US countries. In your assessments of gender, did you account for the extreme population gender imbalance in some of those countries? As merely one example, I see the United Arab Emirates in your study. This country is nearly 70% men.

We thank the Reviewer for raising this concern. From the The World Bank Data on Population, Female (% of total population) data, out of the more than 200 countries, only 5 countries have female population less than 40% (including United Arab Emirates at 30.5%), and only 28 countries have female population less than 49%. Additionally, we found only 2.32% of our selected artists come from these 28 countries. We think this is a small population and it is safe to assume only minor influence on aggregate statistics. Moreover, the art field operates as a global market, hence it is fairly common for artists to exhibit in institutions all around the world, therefore we think it is appropriate to use 50/50 gender portion for the analysis.

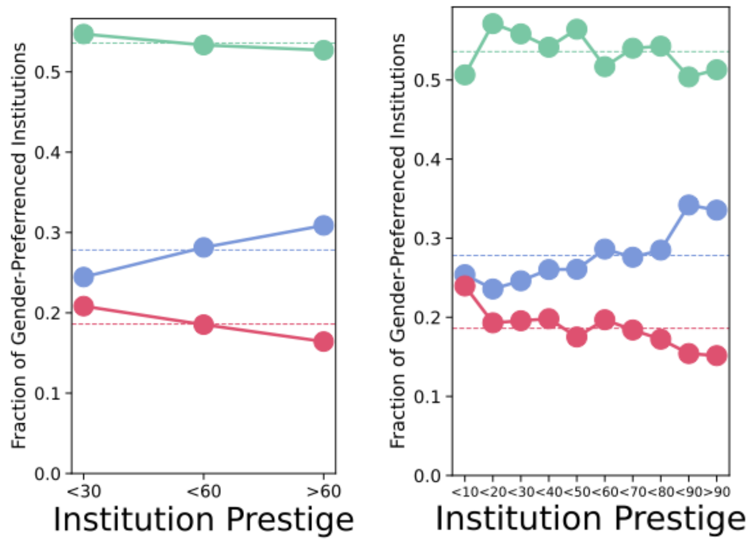
> (3) Another concern relates to your prestige classifications. You state very clearly the cutoffs you use (40th percentile and 70th percentile) to form three levels of prestige, but I would like to see justification for these. In other words, what does the distribution of prestige scores look like? If the prestige scores cluster into three groups, those could be a more natural way to label the institutions. If the distribution is very continuous, I worry that your results might be sensitive to those choices of 40th and 70th percentile, because you are essentially imposing false separation on data that is in reality close together. If you go the route of percentile cutoffs, there should be some sensitivity analysis. In other words, if one changes those cutoffs, do your basic results about the role of prestige stay the same?

This is indeed an important point. We choose to discretize the institution's prestige because in the original paper introducing the prestige score, *Quantifying Reputation and Success in Art,* a similar categorization of artists' career prestige helped establish the lock-in effect. To start, we provide the distribution of institutional prestige (i.e. eigenvector centrality on the co-exhibition network) in the Figure below, which is roughly fat-tailed. As a fat-tailed distribution lacks an internal scale, the choice of where we make any discretization cuts are arbitrary.

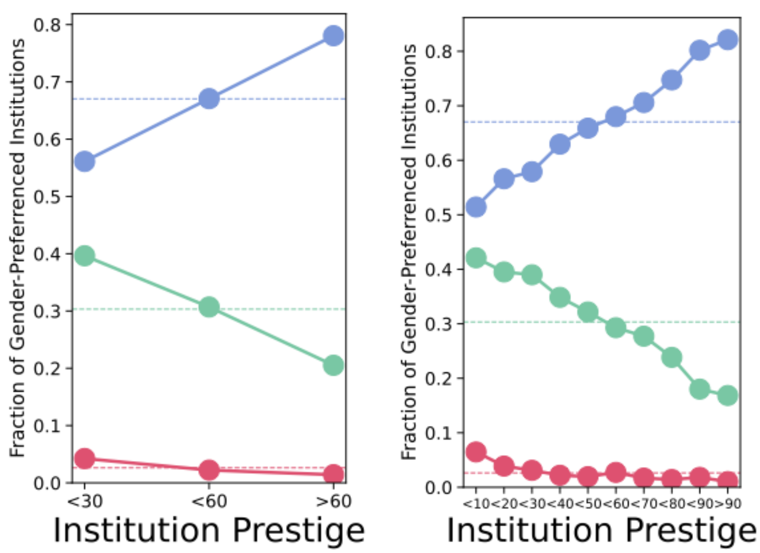


That being said, we completely agree that the robustness of the categorization should be assessed. We therefore replicated our analysis using 1) a different threshold for three categories (at the 30th and 60th percentile), and 2) a decile binning for the institutional prestige. As shown below (and also in the SI, Figure S6A,B) we show the proportion of man-preferred, woman-preferred and gender-neutral/balance institutions for these new prestige categories. The results align with our original conclusion that there are more man-preferred institutions at higher prestige.

Results for gender-neutral using the 30th and 60th percentiles (left) and a decile (right) discretization:



Results for gender-balance using the 30th and 60th percentiles (left) and a decile (right) discretization:



(4) Similarly, you mention that you construct co-exhibition network for each artist with more than 10 exhibitions, but why 10?

To clarify, the institutional co-exhibition network is created using all artist exhibition trajectories. The 10 exhibition threshold was invoked when we classified artists based on the most common co-exhibition gender in their careers. We should emphasize that because the exhibition number is fat-tailed, we are again free to choose any value. Here, we choose 10 as the threshold because we need to compare artists' initial co-exhibition gender to their later co-exhibition gender. Hence, following the calculation of an artist's initial career prestige in Fraiberger et al., 2018 which used the average prestige of their first 5 exhibitions, we set 10 as the lowest number of exhibitions needed to be included in our co-exhibition gender analysis. Following the Reviewer's comment, we now try different thresholds (15 and 20) and find little impact in our conclusion about co-exhibition gender (See SI Section 7C and Fig. S7).

(5) In your logistic regressions of Table 6, I think standard errors should be reported. More substantively, the paragraph at the bottom of page 17 causes me some confusion. On one hand you say you're not interested in predictive ability. On the other hand, you say that exhibitions per year and career length are dominant drivers, and I'd say that it's really hard to assess that if you haven't examined the predictive power of your model. At best, I think you can say that there are positive associations between those two factors and the probability of transitioning. If you want to say they are primary drivers, I think you would need to provide additional evidence. Also, because you fit (I believe) 21 parameters, did you adjust p values for multiple estimates?

We agree with the Reviewer and have now included standard errors in the tables. We want to emphasize that we are not doing causal inference in this model, and to avoid misunderstanding, we revised our language in the main text to clarify this. Finally, all of our p-values are all very small (less than 10^(-10)), so adjusting for 21 estimates would not make a significant difference in our assessment of statistical significance. We now included detailed p-values in the tables.

(6) In the SI, you mention your use of genderize.io. I appreciate the mention of some of the limitation of the algorithm (including around Asian names). Still, I think you should provide more information and possibly some robustness checks. As we both know, genderize provides a probability of the inferred gender, and that probability is actually a frequency calculated as proportion of total occurrences within the underlying database of names. Your choice of a 0.6 probability cutoff feels unjustified to me, and I worry also about the admission of inferences associated with low occurrence counts (you don't mention using a cutoff for these). On one hand, I'd be comfortable with an unjustified cutoff for p if you used a much higher value, but on the other hand, doing so might exclude a lot more data, and possibly not in a gender-unbiased way. In any case, I think it is worth exploring the role that the choice of cutoff plays or somehow providing more justification.

We agree with the Reviewer that all assumptions underlying the inference of gender from first names should be carefully considered, and appreciate the opportunity to address them in greater detail here. Firstly, we make use of the expert curated gender information provided for 207,392 artists (130,848 men and 76,544 women), covering about 42% of the total population. For the remaining 58%, we employ genderize.io, an inference methodology that has been successfully used in many scientific studies and has been shown reliable on large internationally representative samples of individuals (Huang et al., 2020; Kamiri et al., 2016). Following these studies, we selected a confidence level of 0.6. Since we have the expert identified gender information for a large portion of our artists, we compared the genderize.io results (using the original threshold 0.6) with the expert curated gender information. We found that 98.49% of them aligned with the curated gender information.

Following the reviewer's suggestion, we also tried different thresholds, which are shown in the table below.

Threshold	Men Artists	Women Artists	Ratio	Selected Men Artists	Selected Women Artists	Ratio
0.6	58965	27929	2.11	41738	24030	1.736
0.8	56595	26968	2.09	39577	23155	1.709
0.9	55431	26353	2.10	38581	22611	1.706

We see that the overall ratio of men/women artists remains about the same with different thresholds. Though we see a larger difference in the gender population ratio of our selected artists, further analysis shows that the qualitative results on the institution inequality analysis are similar (see below for selected results with 0.9 as the threshold).

Threshold	Men Exhibitions	Women Exhibitions	Exhibition Ratio	Man-preferred Institutions	Woman-preferred Institutions	Gender-neutral Institutions
0.6	662571	355506	1.86	2147	1436	4136
0.9	642011	345311	1.86	2185	1337	4043

We now added these additional analyses to the SI, Section 7A.

Clarity and Context
----

(7) In the abstract, I think the reference to "machine learning" is overblown and/or misleading to most readers. In other words, unless I missed something, you have done logistic regressions. Please don't misinterpret my comment: I think that logistic regressions are a great and appropriate thing to do here. But they certainly existed well before the terminology of machine learning was around. Moreover, I think that in

machine learning, the goal of logistic regression would be classification, and you explicitly say you are NOT trying to build a predictive model. In short, I would remove or at least drastically rephrase this.

We thank the Reviewer for the suggestion and have rephrased it to logistic regression.

(8) General comment for introduction: The audience of this journal might benefit from an explanation of the difference between a museum's permanent collection and its exhibitions. There are a lot of basic things about how art museums work that a general science and/or social sciences audiences will not know.

We thank the Reviewer for the suggestion and have added some background information about the differences between a museum's permanent collection and its exhibitions to the Introduction.

(9) On page 1, I might remove the word "quality" in your phrase about "nature of the quality of the art." For a gender disparity to be rooted in quality of the art requires that quality of the art is a thing that objectively exists, which I am not certain I believe.

We thank the Reviewer for the suggestion and have rephrased that section.

(10) On page 3 and throughout, I applaud you for addressing the limitations of using a binary coding of gender. Still, I would recommend going a step further and explicitly addressing/admitting that this may result in misgendering of individuals and/or erasing identities. To be clear, there's no way to do this research WITHOUT harm/erasure, and I am not suggesting that the research not be done. You could use the limitations of the data as a call for more inclusive gender data. You may have said this once but I think saying it right from the start of the paper is both appropriate (which is most important) and might protect you from visceral pushback (which is a secondary benefit).

We thank the Reviewer for the suggestion and we have added more discussion about the potential harm of using binary coding for gender, and also emphasized the importance of more inclusive data.

(11) On page 4, when I got to this section, I was genuinely confused because I did yet not know what was to come later. As I read the main two paragraphs, I kept jotting notes like "this strikes me as the wrong comparison," and "why no statistical test?" At the end of the section you state that the goal is to move beyond these simple measures, but I wonder if your readers would be better served by putting that up front: "we're going to do something ludicrously simple and then we're going to move on to something more sophisticated."

We thank the Reviewer for identifying this point of confusion. We have now revised the manuscript to clarify the presentation of baseline assumptions which empower our more sophisticated analysis.

(12) On page 8, you begin talking about prestige, but you only define the way prestige is calculated later on. I think this might be confusing for many readers and I would recommend rearranging or at the very least, foreshadowing.

We thank the Reviewer for catching this problem. And we have now rephrased the initial mention to the word prominent since it only refers to an exemplar list of institutions.

(13) Also on page 8, you use the abbreviations "MoMA" and "MUMOK." Most readers will not know these, so you should define them. This also holds for the caption of Figure 2, where you mention MoMA.

We thank the Reviewer for this suggestion, and have included the full name of the institutions.

(14) On page 9, in Figure 2, the horizontal axis labels in panel (e) run together and should be fixed.

We thank the Reviewer for catching this problem. We have now updated the Figure.

(15) In the SI, on page 12, writing "0.00" for p values is not helpful. Can you please give more of a sense of the order of magnitude, or a coding with asterisks for different levels of significance, or anything else?

We thank the Reviewer for the suggestion, and we have now included more detailed p-value information.

References

The references are appropriate.

In summary, we thank the Reviewer for their strong endorsement and detailed suggestions, which greatly improve our manuscript. Following their excellent recommendations, we have conducted robustness checks on several of our data processing decisions and statistical tests, all of which further validate our analysis of influences to gender inequalities in the art world.

Reviewer #3 (Remarks to the Author):

Summary: The authors consider the gender inequality in contemporary visual arts. In particular, authors distinguish two forms: gender balance (population proportions) and gender neutrality (Difference between the visibilities in exhibitions of those already in the populations). These metrics are also dissected based on institutional prestige to depict the intersectionality of institutional prestige and gender inequality.

Although gender inequalities have been explored in various fields and communities, this is the first effort that I am aware of that extended and effort to do so for the field of visual arts. Additionally, the datasets that the authors have made available will hopefully lead to more efforts continue this area of work. In that sense, I appreciate the authors' effort. However, I have several comments about this work which I elaborate below:

We thank the Reviewer for their detailed assessment of our project and acknowledging its novelty.

Comments:
1. Several previous works focussed on gender inequities in citations using very similar notions that the authors use (gender balance and gender neutrality). They also have have reached conclusions that agree with and strengthens the conclusions that the authors reach. In particular, the following three papers (in chronological order) are heavily related to the authors' work in my opinion:

(i) Dworkin, Jordan D., Kristin A. Linn, Erin G. Teich, Perry Zurn, Russell T. Shinohara, and Danielle S. Bassett. "The extent and drivers of gender imbalance in neuroscience reference lists." Nature neuroscience 23, no. 8 (2020): 918-926.

(ii) Nettasinghe, Buddhika, Nazanin Alipourfard, Vikram Krishnamurthy, and Kristina Lerman. "Emergence of structural inequalities in scientific citation networks." arXiv preprint arXiv:2103.10944 (2021).

(ii) Teich, Erin G., Jason Z. Kim, Christopher W. Lynn, Samantha C. Simon, Andrei A. Klishin, Karol P. Szymula, Pragya Srivastava et al. "Citation inequity and gendered citation practices in contemporary physics." Nature Physics 18, no. 10 (2022): 1161-1170.

(i) mentions "…..we find that reference lists tend to include more papers with men as first and last author than would be expected if gender were unrelated to referencing."
(ii) mentions "…female authors, who represent a minority of researchers, receive less recognition for their work (through citations) relative to male authors; second, authors affiliated with top-ranked institutions, who are also a minority, receive substantially more

recognition compared to other authors…..we find that merely increasing the minority group size does little to narrow the disparities….”
(iIi) mentions “citation imbalance in favor of man-authored papers is highest for papers authored by men, papers published in general physics journals and papers for which citing authors probably have less domain or author familiarity.”

The above findings (in particular from ii and iii) agrees with what authors find: merely getting rid of gender balance may not be sufficient to achieve gender neutrality (they have used definitions which normalizes for one factor while considering the other). Additionally, they also suggest strategies to mitigate the gender inequality via strategies that compliment the authors’ strategies. As such, I believe a discussion of such highly related work (though not on the field of visual arts) and how they align/differ from the authors’ conclusions will be a good addition to the paper.

We thank the reviewer for pointing us towards some excellent discussions of gender inequalities in science by the Bassett and Lerman labs, and completely agree that the scientific and artistic domains have much to learn from each other. Following the Reviewer’s recommendation, we have expanded the discussion of quantitative methods for understanding scientific gender inequality in our introduction and added these pertinent references.

2. The authors use the term “systemic gender inequality”. However, most of the authors’ analysis is descriptive in nature (via comparison of proportions, p-values, etc.) and the model they consider is probabilistic rather than causal. Could the authors shed more details as to how they justify the observed disparities being “systemic”? For example, the values in Fig. 1(a) for population and exhibition categories agree very closely with each other leading to the question “is there a causal relation between gender balance and gender neutrality and gender disparity in auctions?” If so, in which direction? I do not have an answer to offer here. But discussing these aspects (or at least highlighting the difficulties to explore them) will justify the observed disparity being “systemic”.

To clarify, the term “systemic gender inequality” refers to those aspects of inequality which extend beyond the decisions made by any one institution and instead manifest as collective consequences of the system. In this respect, an artist’s co-exhibition gender arises from the gender preferences of all institutions in which they exhibited throughout their career, and gender preference assortativity in the co-exhibition network arises from institutions exhibiting similar populations of artists. Following the Reviewer’s comment, we have expanded this discussion in the main text and further highlighted that our work is not casual.

3. The authors have considered the network of institutes. Is there an analogous form of a network of artists which will reveal another aspect of the picture? E.g., in terms of the structure, do these two networks differ? If so, do those differences shed more light on the? This question occurred to me as a recent work which I came across (Lerman, Kristina, Yulin Yu, Fred Morstatter, and Jay Pujara. "Gendered citation patterns among

the scientific elite." Proceedings of the National Academy of Sciences 119, no. 40 (2022): e2206070119.) used the difference between ego networks of authors to classify their gender, the point being that there is such a larger difference between networks of men and women so that an ML algorithm can use them to identify the gender. Maybe this could be helpful (at least for future work).

We thank the Reviewer for directing us towards interesting questions for future work. While the co-artist network (a projection of the artist-institute exhibition career network onto the artist nodes) has been a subject of study in our lab for a while, we have not yet found a way to tease insights about inequality in artistic careers. No doubt, as other great works such as the mentioned Lerman paper help to refine and develop techniques for ego networks, we will pursue additional research in that direction.

In summary, we thank the Reviewer for their literature recommendations and detailed questions, all of which helped to enhance the clarity of our work.

Reviewer #4 (Remarks to the Author):

This manuscript aims to understand the gender disparities in the world of contemporary visual arts. They found systematic gender differences in terms of artists, exhibition opportunities, auction opportunities, and auction prices. I found this study interesting and does provide first-hand evidence of the gender disparity issues in visual arts.

We thank the Reviewer for acknowledging the value of our work for revealing quantitative trends in the gender inequality issues in visual arts.

One concern for the analysis was that the artist productivity factor was not mentioned anywhere in the study, which could possibly interfere with the number of exhibits and auctions.

We thank the Reviewer for this comment. We fully agree that it would be interesting to analyze how an artist's productivity is related to their success. Indeed, in science productivity is a key variable in the analysis of scientific careers. Sadly, our data only captures presence in exhibitions and auctions, and does not contain information on the specific artworks created by the artists. To the best of our knowledge, only the "catalog resume", typically assembled after the artists' death, can capture true productivity. Moreover, exhibition and auction sales don't necessarily imply a new artwork was created, displayed or sold, as it is fairly common to have the same art piece being exhibited or sold multiple times. We now discuss this limitation in the Introduction and SI, Section 1A.

Page 2: I also found this argument problematic. The manuscript states, "Studies show that people cannot infer the gender of the artist from the work itself [5, 2], suggesting that the observed gender disparity cannot be rooted in the nature or the quality of the art men and women produce. As Nochlin hypothesized, the answer is likely rooted in 'the very nature of our institutional structures.'". My understanding here is that if, as claimed by the manuscript, people are unable to infer the gender of the artists, then why would it suggest "that the observed gender disparity cannot be rooted in the nature or the quality of the art men and women produce"? I had a hard time making the connection.

We have now rewritten the discussion of the previous literature to highlight that the question of whether artistic quality is agnostic to gender has not been resolved, and indeed, that the empirical evidence is mixed. Regardless of whether the artwork itself differs between men and women artists, our work demonstrates that social forces are highly correlated with gender equity in art, and helps reveal the emergent systemic signals of gender inequality.

Figure 4: I found it hard to reach the conclusions the manuscript claimed based on Figure 4a based on direct observation. (1) What color was used for the edge connecting a man-preferred and woman-preferred institution? The choice would have an impact on the visual impression of the network. (2) Nodes in the more peripheral locations do seem

to show inequality-based assortativity, but it is hard to tell about the large proportion of the nodes in the center of the network. I think the reader would benefit from additional indicators or different network representations to support the claim.

We thank the Reviewer for their critical assessment of the co-exhibition gender assortativity. Our aim of Figure 4 is to provide a visual representation of the network and the institution inequality, which we believe can help the reader understand more about how the nontrivial structures hidden in artist careers bias access to future opportunities. Rather than drawing the inequality-based assortativity conclusion directly from eyeballing, we do quantify it, an analysis that we originally included in SI Section 3C. However, following this review, we have now moved and expanded this analysis within the main text.

Specifically, to reveal neighborhood bias, we calculate the percentage of weighted outgoing links to institutions with each gender representation inequality under the gender-neutral criteria. We find that man-preferred institutions connect to other man-preferred institutions much more than would be expected by chance, 34.1% compared to the random baseline of 27.77%. Similarly, woman-preferred institutions connect to other woman-preferred institutions much more than would be expected by chance, 23.5% compared to the random baseline of 18.57%. We now include these numbers in the main text as well.

Besides the neighborhood computation, we also calculated the assortativity coefficient (Newman, 2003). We found that our network shows an assortativity coefficient of 0.020, with standard deviation 1.22e-07, while the random shuffled network has a coefficient of -0.0006 with standard deviation 1.20e-07.

Additionally, we conducted an analysis of multi-scaling assortativity introduced by Peel et. al. (2018) *PNAS*, which provides more fine-grained observations on network assortativity. Since we are mostly interested in the potential for gender-preference to influence artist career trajectories, we calculate the multi-scaling assortativity focusing on man-preferred and woman-preferred institutions, (under the gender-neutral criteria). As seen in Figure A, below, there are two peaks in the multi-scale assortativity, both of which occur at larger values of assortativity than expected under the null model. Separating on the gender preference, we can find that the man-preferred institutions show stronger assortativity than the woman-preferred institutions, but both classes are considerably more interconnected than expected by chance.

One issue with both aforementioned measures of assortativity is that they do not account for the varying edge weights in our network, and instead their application forces us to make the assumption that all institutional interactions are equal. Therefore, both measures underestimate the extent of assortativity in this network. To address this challenge, we conduct another analysis in which we filter the network based on the edge confidence score. The confidence score is a statistical measure reflecting the likelihood for a specific transition between institutions to have occurred by chance based on the volume of exhibitions at both the source and target institution. Therefore, the filtered network represents the network backbone of significant transitions between institutions. As shown in Figure B, below, applying the multi-scaling assortativity on this filtered network reveals even higher levels of gender-preference assortativity. The effect is exceptionally strong for man-preferred institutions which significantly differ from the expected assortativity under the null-model.

We have now revised the main text and SI with an extended discussion of the gender-preference assortativity analysis.

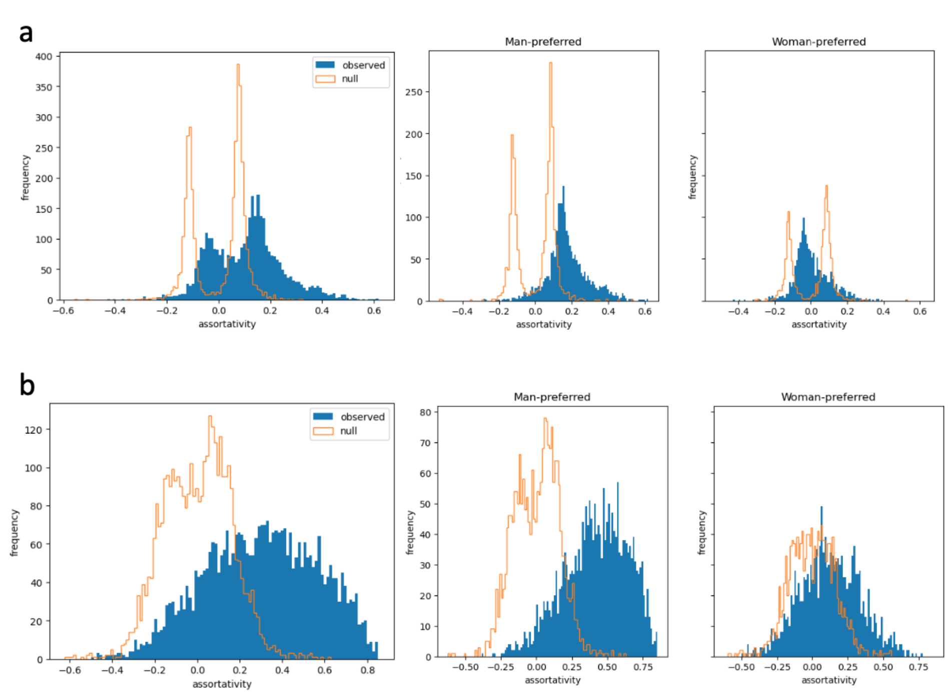


Figure 5 needs legends.

We thank the Reviewer for catching this problem. We now update the legends for Figure 5.

In summary, we thank the Reviewer for their detailed comments and recommendations, all of which have greatly helped to enhance the clarity of our manuscript.

Reviewer #5 (Remarks to the Author):

Comments on:

Quantifying Systemic Gender Inequality in Contemporary Visual Art

Overview

While the most interesting scientific papers clearly identify a hypothesis to be tested, typically derived from a particular theoretical framework, and proceed to test the hypothesis using observed data, there are potential contributions to be made also by descriptive papers that measure a phenomenon or relationship with care and present new and unexpected findings that can help encourage development of theory or identification of hypotheses to be tested. This paper could potentially be put into this second category. Providing, as suggested by the title, a quantitative measurement of gender inequality in contemporary visual art. Presumably this would mean the art market, consisting of the production, consumption, sale and final prices of art works.

The central findings reported in this paper are:

1. Combining both museums and commercial galleries into the category of “art institutions” almost 54% are gender neutral (exhibiting artists in proportion to the gender share in the population of contemporary artists). About 30% are gender balanced (half of artists exhibited from each gender).
2. Among those that are not gender-neutral, museums are split almost evenly between favoring women versus favoring men, while commercial galleries split almost 2 to 1 favoring men over women artists.
3. Using a notion of “prestige” for exhibition venues determined in an earlier paper, higher prestige institutions tend to favor male artists. This difference is very small based on gender neutrality, but larger based on gender balance.
4. Of the 65,768 artists analyzed, 24% of the men and nearly 16% of the women had at least one auction sale. In terms of total value of auction sales, the value of works sold by men artists was over 7.5 times that of women artists although most of this difference is due to the number of objects sold rather than the average price of the objects. The average sales price of an object sold by men was 1.75 times the average price of a work by a woman artist.
5. A higher proportion of museums are gender-neutral than of commercial galleries (60% versus 50%), although it should be noted that over 60% of art institutions, accounting for 11% of exhibitions have too few exhibitions to be reliably categorized.

Somewhat less successfully, the paper puts forward some additional claims. I will explain below what I see as problems with these results, but I list them here because the authors put them forward as central findings.

6. Art institutions with comparable degrees of gender inequality tend to be connected in a directed exhibition network to each other, often exhibiting the same artists.
7. Artists (whether male or female) may have careers that result in their exhibitions being at predominantly man-preferred, woman-preferred, or neutral exhibition venues. Comparing an artist's first five exhibitions to their last five exhibitions reveals considerable stability to this exhibition proclivity, and the tendency to stay in the same category is somewhat stronger for artists who exhibit at low-prestige venues.

Novelty of results and review of related scholarship

The problem of gender inequality in the art world is an issue that has received considerable attention from other scholars whose results have been published previously. Some of these are cited as old versions and should more properly be cited as the versions that have passed peer-review. This includes Adams, et al. (2021), Bocart, et al. (2021) and Cameron, et al. (2019) (full citations given below). Other relevant and important contributions are not cited at all, for example Cowen (1996), Droege (2022), Galenson (2007), Greenwald (2021), Hoffman & Coate (2020), and LeBlanc & Sheppard (2022).

We thank the Reviewer for suggesting connections to relevant economics scholarship, and appreciate the opportunity to situate our contribution within this literature. We have now updated our literature review to engage with this fascinating scholarship.

There are occasions in the paper when citations are offered that seem to betray a misunderstanding of the contents of the cited work, or where the cited work seems inappropriate. For example, at the bottom of page 4 the manuscript asserts that pressure to address gender imbalance "… stems from the strong belief and evidence that artistic performance and artistic aptitude are agnostic to gender." A citation is given to Bourdieu and to the work of Adams, et al. (2021). Bourdieu's beliefs are often inchoate and vague, and certainly his 1993 collection of essays presents almost no evidence or formal analysis concerning gender and artistic skill1. Beliefs could be cited on both sides of the issue (one might as well include Georg Baselitz's "belief" that women don't paint very well).

Adams and her collaborators do address this issue, but come to a different conclusion from the one asserted. To be "agnostic to gender" would imply that works of art convey no information about the gender of the artist who created the work. Yet Table 4 and Figure 1 in Adams, et al. (2021) show that using titles as indicators of image content, they can differentiate between artworks produced by men and by women. In the final section of their paper, they do show that an online survey of small to modest samples representative of the general US population are mostly unable to reliably determine the artist gender based on online images. They note that the results won't generalize if this sample is not representative of art collectors or art auction participants, and their conclusion is that for this sample representative of the entire US population, "… structural differences that might exist are not readily observable."

Differences between artworks by men and by women are also noted in the paper by LeBlanc and Sheppard (2022). Their analysis shows that women artists produce works that are smaller (about 16% smaller in area) and less chromatically diverse images, that are less likely to be portraits or landscapes, or dated, or produced using oil paints on canvas. Interestingly, their analysis uses the Google SafeSearch image analysis neural network to show that women artists are more likely to produce images classified as likely to contain Adult or Racy images.

So while there is "belief" existing on both sides of the issue of whether artworks are "agnostic" to gender, the actual empirical evidence is at best mixed and in my view actually support the assertion that art works by women DO differ in some important ways from works by men. A more careful and complete reading of published research on the subject would improve the exposition of this paper.

We thank the Reviewer for their insights and literature overview regarding "whether art is agnostic to gender". We realize this issue is quite complex and multifaceted. However, regardless of whether the artwork itself differs between men and women artists, our work demonstrates that social forces are highly correlated with gender equity in art.

As suggested by both the works cited in this paper and the additional works mentioned above, the problem of under-representation of women artists in exhibitions, auctions, and sales is widely recognized and has been the subject of scholarly inquiry for over 50 years. The findings of this paper that provide quantitative confirmation of this are therefore not really novel or surprising. If anything the gender gaps that are identified in this paper are somewhat smaller than might have been expected given the large volume of scholarship as well as journalism and commentary on the topic. This makes central findings 1 through 4 listed above useful confirmation, but not a significant contribution to the literature. This literature, indeed, has been "moving on" to more carefully test hypotheses about why these gender gaps exist. Thus for example Adams, et al. (2021) test the importance of local social attitudes towards women as an important determinant of the gender gap. LeBlanc and Sheppard (2022) consider the intersection of gender and ethnicity and different characteristics of the works themselves. Greenwald (2021) by contrast, looks at supply constraints arising from limited access of women artists to formal training or the demands of family and personal life as a factor.

We thank the Reviewer for providing another opportunity for us to highlight our study's primary contributions to understanding the scope of gender inequality in art. To the best of our knowledge, and further supported by the Reviewer's literature suggestions, our data provides the most comprehensive window into gender inequalities in art, being unique in combining a global coverage, with detailed exhibition information from both museums and galleries, and auction sales. This data empowers our work to juxtapose the extent of gender inequity in different countries, institutional types, and prestige. Indeed, three of the works highlight by the Reviewer: Adams et al. (2021), LeBlanc & Sheppard (2022), and Bocart et al. (2022) focus only on auction sales, artist demographics, and artwork characteristics, but do not integrate this with exhibition data and are thus unable to draw insights into how prestige and network effects influence gender inequalities in art. Since there is nuanced and complex interplay between artistic production, museum exhibition, gallery exhibition, and auction sales, our study constitutes a new direction in the study of gender inequality in art.

> The current manuscript has one result that seems as if it could provide an avenue for a test that has not been emphasized by other scholars: the difference in gender gap between museums and commercial galleries listed above as central finding 5. This is novel in part because very few scholarly analyses have access to data on outcomes in commercial galleries, relying almost exclusively on art auction sales instead. Galleries are businesses who, in pursuit of their objective to earn profits for their owners, must cater to the desires of collectors and others who are their customers. Museums, by contrast, are in almost all cases government or non-profit organizations with more complicated missions to preserve and care for objects, to attract and educate the public, while still generating sufficient revenues to “keep the doors open”. While not conclusive, finding 5 might be generally interpreted as supporting the idea that the gender gaps observed in the art market are the product of “demand side” preferences of collectors and buyers rather than supply constraints.

We agree with the Reviewer that it would be very interesting to learn if gender inequalities in art are driven more by demand side preferences of collectors and buyers or by supply side constraints of artists. Answering this question requires, however, the data from the demand side (namely collectors and buyers), which is never disclosed by galleries. Indeed, despite our best attempts at scraping together a representative dataset in this direction, we have found that the identity of both curators and collectors are kept mostly private. However, we agree this would be a great direction for future research should the data become available, and prompted by the Reviewer’s comment, we include it into the discussion.

> Unfortunately, the paper cannot or does not test this potential hypotheses, choosing instead to regard galleries and museums as essentially equivalent. Virtually every mention of one type of institution, in both the main article text and the supplemental material, is within one word of the other type separated only by the word “and”. While the general public may make this mistake, it really is inexcusable for scholars of the art market to conflate the two. While both of these institutions stage exhibitions and artists are anxious to be included, listing exhibitions of both types on their CVs or web sites, they are not equivalent. In general, artists would prefer to have shows at museums and would prefer for their work to be sold to museums. I have personally had the experience of working with museums seeking to acquire the work of an artist from a gallery and, once it is know that the sale would be for a museum collection an immediate discount in price of 10% is offered. To regard the two types of institutions as equivalent is a

fundamental error in understanding the art market and misses a real opportunity to illuminate a possible source of the gender gap.

We thank the Reviewer for their unique insights into the art world. However, rather than treat all institutions as equivalent, our work seeks to quantitatively differentiate the prestige of the institution and therefore its influences on artists' careers and consequences for gender equity. One feature of this, which mirrors the Reviewer's intuition, is that the average prestige of museums is higher than galleries, reflecting their more visible place in the art world. However, we further argue that not all galleries or museums are the same, and that, for example, the Gagosian Gallery in NYC is far more prestigious than the Palm Springs Art Museum in California. While in our analysis we label them differently, we study them as part of the same network, for the very reason the Reviewer mentions—they are both integral to the same art ecosystem.

Network and data analysis methodology

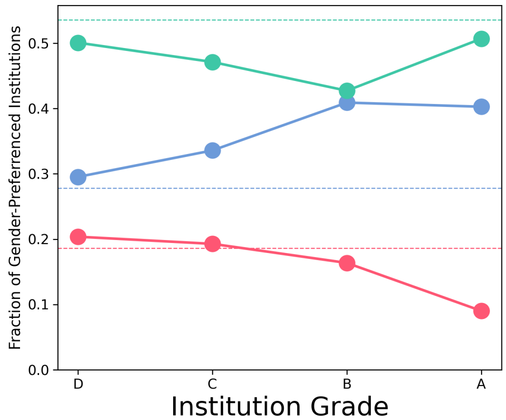
This fundamental asymmetry between the types of institutions included in the sample raises further difficulties for the network analysis and measurement of "institutional prestige" that is offered as an important variable in understanding the exhibition gap between men and women artists. The first problem that arises is that with exhibition venues as the nodes and the strength of links or flows between nodes determined by the number of artists exhibiting first in one venue then in the other, the intrinsic difference in the institutions means that a link from a gallery to a museum should not be regarded as making the same contribution to institution prestige as a link from a museum to a gallery. The former (say an exhibit at Hauser & Wirth followed a few years later by a show at the Tate Modern) is a signal of a successful artist whose career is building. By contrast the latter type of link (say an exhibit at Marian Goodman or Paula Cooper a few years after the Tate exhibition) may not say much about the artist, but more about the gallery taking advantage of the museum exhibition to generate profitable sales to collectors seeking to acquire a blue chip artist's work. These nuances are missed by the network methodology as it is applied.

We thank the Reviewer for their careful consideration of the network construction used in our formulation of prestige. We agree that there are many important differences between galleries and museums. However, every modeling problem requires the researcher to make decisions about which details to include in the model. In our case, we applied the framework developed in *Quantifying Reputation and Success in Art* to quantify the prestige of institutions based on the co-exhibition network. In that paper, they demonstrated explicitly that simplifying artistic careers by considering galleries and museums as equivalent is sufficient to accurately predict institutional rankings derived from both expert-curated and total auction sales. One emergent consequence of this formulation is that, on average, museums have higher prestige than galleries, mirroring the Reviewer's intuition. As this analysis was already conducted in Fraiberger et al. 2018, we did not feel the need to repeat it here.

A related difficulty arises when, as in the results presented in Figuires 2c and 2d and discussed on page 10, the authors try to use the “prestige” of the exhibition venues as an explanatory variable for the fraction of venues that are gender neutral or gender balanced. The network of art exhibition venues with links defined as the authors have done presents a classic example of a network with endogenous link strength. Such networks have been extensively researched in the economics literature and the authors might benefit from a careful study of papers such as Bloch & Dutta (2009) or Olaizola & Valenciano (2020).

Of course, if the link strength is endogenous then the eigenvalue centrality will also be endogenously determined by the choices and investments made by the individuals (artists or gallery owners) and organizations (museums) that comprise the network. I’m sure I don’t have to explain to the data scientists among the list of authors that this endogeneity creates potential statistical bias in estimating or displaying the relationship between the “prestige” as measured by network centrality and a potentially noisy outcome like the fraction of venues that are gender neutral. While the existing paper does not emphasize estimates of parameter values using prestige rank as an explanatory variable, the presentation and comparison of slopes in Figures 2c and 2d are essentially the same thing and subject to the same problems of bias due to endogeneity.

We thank the Reviewer for raising this valid concern. As the prestige and the gender preference assignment are all related to artists’ exhibition trajectory, there is indeed a possibility that we have a problem of endogeneity. We first note that our conceptualization of institutional prestige is grounded in expert assessments for a subset of institutions—our network centrality based prestige measure only serves as a method to extend the expert curated assessments to all institutions in our data. Prompted by the Reviewer’s observation, we decided to eliminate the potential for endogenous variables in our model by replicating the analysis using only the subset of institutions for which we have an expert’s assessment of institutional prestige (graded such that A reflects the most prestigious). We found that the extent of inequality at the most prestigious institutions is magnified: there are more man-preferred institutions and less woman-preferred institutions at higher grades. We have now added this analysis in the SI, Section 7B and SI Figure 6.

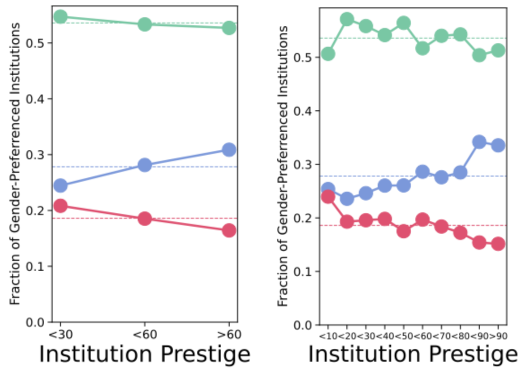


Increasing proportion of co-exhibition man institutions at high prestige using the expert assessment of institution prestige (A is most prestigious).

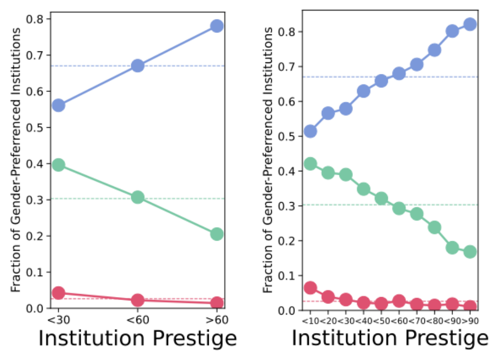
> I also that it curious that the authors grouped exhibition venues together into “low”, “mid” and “high” prestige groupings with dividing lines at the 40th and 70th percentiles. First one might ask why group them at all? This is not done, for example, in the 2018 Science companion paper (see for example Figure 2B on page 827).
>
> Grouping them has the effect of suppressing the ability of the reader to perceive noise in the data that can help evaluate the strength of the purported relationship. The situation is not helped by choosing unorthodox percentiles to create the groups and not offering explanation. Why not create the groups based on terciles of the prestige ranking data?

To clarify, we choose to discretize the institution’s prestige following the original paper introducing the prestige score, *Quantifying Reputation and Success in Art,* which used such binning to fruitfully categorize artists’ career prestige to establish the lock-in effect (see Figures 3 and 4). To test the robustness of choosing the cutoff threshold for the categorization, we now tested 1) different thresholds for three categories (at 30th and 60th percentile), and 2) categorization into 10 categories (at each percentile) for the institutional prestige. Below we show the proportion of man-preferred, woman-preferred and gender-neutral/balanced institutions at different prestige categories, and found the consistent conclusion that there are more man-preferred institutions at higher prestige.

Results for gender-neutral using the 30th and 60th percentiles (left) and a decile (right) discretization:



Results for gender-balance using the 30th and 60th percentiles (left) and a decile (right) discretization:

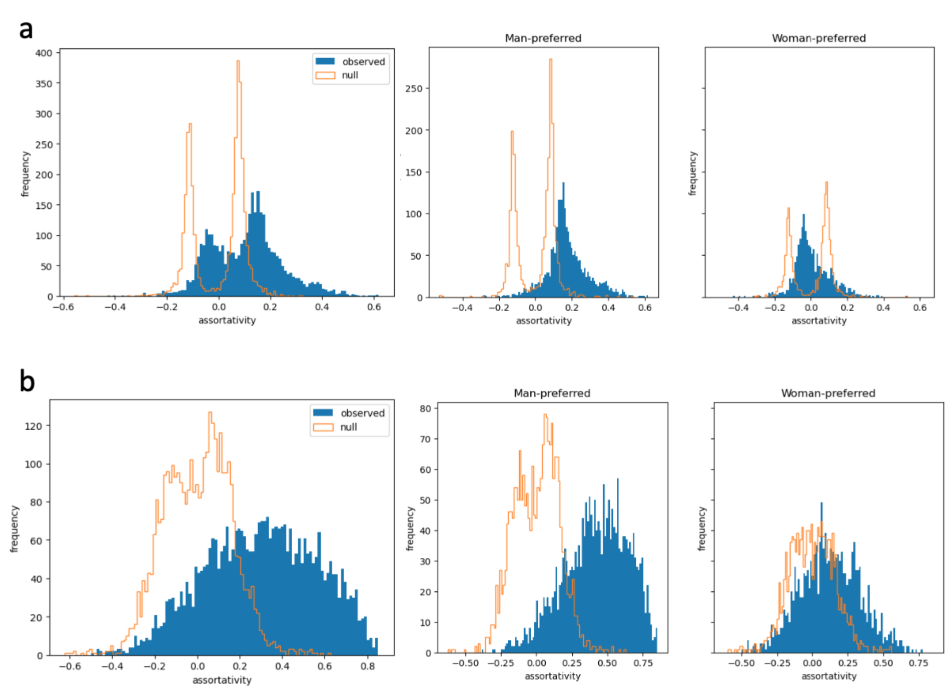


Finally, I found the presentation of the Co-exhibition networks in Figure 4, along with the claim on page 11 that the apparent "… emergence of clusters dense in blue or red nodes is evidence of inequality-based assortativity, indicating that institutions with comparable degrees of inequality tend to be connected to each other, often exhibiting the same artists" to be potentially misleading and inaccurate. Simply looking at the Figure the reader might reasonably get the impression of a random assemblage of red and blue dots and lines. We are told on page 10 that the network is printed using a force-directed algorithm, but no other details are provided. Producing meaningful printouts of complex networks is a notoriously difficult problem, and while force-directed algorithms of one sort or another are generally used, there are several specifications for the attractive forces and repulsive forces for the nodes that can affect what clusters appear to the eye of the beholder. To assert that there is an "emergence of clusters" feels like the authors trying to use a "Jedi mind trick" to convince the reader that there is something present that they are unable to establish through ordinary data analysis.

We thank the Reviewer for highlighting the importance of the network gender preference assortativity analysis. Following their excellent recommendation, we have now updated the text describing Figure 4 to better clarify the network construction and visualization.

We also want to emphasize that the network was displayed only for illustrative purposes and that our main conclusions were drawn from a separate assortativity analysis previously discussed only in the SI (Section 3C). To avoid further confusion, we have now moved and expanded the quantitative analysis of gender preference assortativity to the main text. Specifically, to reveal neighborhood bias, we calculate the percentage of weighted outgoing links to institutions with each gender preference under the gender-neutral criteria. We find that man-preferred institutions connect to other man-preferred institutions more often than would be expected by chance, 34.1% compared to the random baseline of 27.77%, which are statistically different with a p-value of less than 10^(-102). Similarly, woman-preferred institutions connect to other woman-preferred institutions more often than would be expected by chance, 23.5% compared to the random baseline of 18.57% which are statistically different with a p-value of less than 10^(-87). Besides the neighborhood computation, we also calculated the assortativity coefficient (Newman, 2003). We found that our network shows an assortativity coefficient of 0.020, with standard deviation 1.22e-07, while the random shuffled network has a coefficient of -0.0006 with standard deviation 1.20e-07.

Additionally, we conducted an analysis of multi-scaling assortativity introduced by Peel et. al. (2018) *PNAS*, which provides more fine-grained observations on network assortativity. Since we are mostly interested in the potential for gender-preference to influence artist career trajectories,man-preferred and woman-preferred institutions, we calculate the multi-scaling assortativity focusing on man-preferred and woman-preferred institutions,these two types of institutions (under the gender-neutral criteria). As seen in Figure A, below, there are two peaks in the multi-scale assortativity, both of which occur at larger values of assortativity than expected under the null model. Separating on the gender preference, we can find that the man-preferred institutions show stronger assortativity than the woman-preferred institutions, but both classes are considerably more interconnected than expected by chance.



One issue with both aforementioned measures of assortativity is that they do not account for the varying edge weights in our network, and instead their application forces us to make the assumption that all institutional interactions are equal. Therefore, both measures underestimate the extent of assortativity in this network. To address this challenge, we conduct another analysis in which we filter the network based on the edge confidence score. The confidence score is a statistical measure reflecting the likelihood for a specific transition between institutions to have occurred by chance based on the volume of exhibitions at both the source and target institution. Therefore, the filtered network represents the network backbone of significant transitions between institutions. As shown in Figure B, above, applying the multi-scaling assortativity on this filtered network reveals even higher levels of gender-preference assortativity. The effect is exceptionally strong for man-preferred institutions which significantly differ from the expected assortativity under the null-model.

We have now revised the main text and SI Section 3C with an extended discussion of the co-exhibition gender assortativity analysis.

> I did find the exploration of the idea of defining the co-exhibition gender of artists, and the potential change of co-exhibition gender, potentially interesting. There is again the problem of endogeneity of the network that is used as the basis of the presentation in Figure 5d and the inference that there is a difference in co-exhibition transition probabilities depending on low, mid, or high prestige exhibition venues, but the idea is novel and potentially interesting. A main difficulty is that it is not clearly whether the co-exhibition gender is related to an outcome that we really care about. It is not clearly related to actual gender of the artist (a male artists could spend his entire career as a

co-exhibition female, or a woman artist her entire career as a co-exhibition male) and therefore not related in any simple way to the gender gap in exhibition frequency.

We thank the Reviewer for providing another opportunity for us to elaborate on the importance of co-exhibition gender on artist careers. Our final model demonstrates that the co-exhibition gender is more predictive than traditional gender when modeling access to the auction market, thus we argue it is capturing an important feature of the art ecosystem. In this sense, the manuscript argues that gender gaps in art manifest themselves through differences in: 1) exhibition frequency, 2) exhibition prestige, and 3) co-exhibition gender. Our analysis further reveals that not all exhibitions at the same prestige are equivalent, but differ based on perceived biases against the exhibiting community. The consequence for individual artists is that access to the auction market is influenced by the composition of the co-exhibiting community, regardless of their actual gender. However, we also agree with the Reviewer that the analysis cannot establish causality, and the inequality may be rooted in other facets of the art world that are correlated with co-exhibition gender. Still, we feel this analysis does reveal emergent systemic features of gender inequality in art.

We have now expanded our discussion section to provide additional elaboration on the interpretation and importance of co-exhibition gender. Surprisingly, our analysis reveals that there are nearly equal numbers of co-exhibition man and co-exhibition woman artists (see the Table below), and that women are much more likely to be categorized as co-exhibition women, while most co-exhibition men are also men.

Gap in co-exhibition gender

Co-exhibition gender	**Co-exhibition Neutral**	**Co-exhibition Man**	**Co-exhibition Woman**
Number of Artists	6387	11007	11327
Number of Men Artists	4169	8947	5154
Number of Women Artists	2218	2060	6173
Mean of Exhibition Count	24.80	35.19	29.50
Number of Artists with Auction	1428	4295	2340
Portion of Low Prestige	29.14%	25.86%	27.95%
Portion of Mid Prestige	40.25%	37.43%	41.50%
Portion of High Prestige	30.61%	36.71%	30.55%

The analysis presented suggests that co-exhibition gender, or interaction between actual artist gender and co-exhibition gender, may play a role in affecting the probability that an artist ever has a work offered for sale at auction. Unfortunately the failure to include any other controls (such as medium, support, or image content) as factors in determining appearance at auction, or the failure to explore the relationship between these variables and actual auction price make it difficult to tell if we (or if artists) should really care about this variable. Still, I would encourage the authors to continue to explore ways in which this measure might be a useful metric for understanding the importance of gender in the art market.

We thank the Reviewer for raising concern of other possible factors that may contribute to an artist's access to auction. Following their excellent recommendation, we have added two additional models to our manuscript:, Model (7) adds the proportion of man-preferred and woman-preferred exhibitions throughout an artist's career, and Model (8) adds the primary medium of each artist's body of work. We found that the general trend that co-exhibition gender has higher importance than artists gender still holds, although additional comparisons are hampered by the fact that medium information is not provided for a large portion of our population. We have now included these results in SI 6B.

Model		**(7)**				**(8)**			
Number of Observations		16126				1781			
BIC		19616.37				1983.25			
Variables		Coef.	O.R.	S.E.	P Val.	Coef.	O.R.	S.E.	P Val.
Intercept		-4.97*	0.01	0.14	$<10^{-10}$	-4.31*	0.01	0.53	$<10^{-10}$
Exhibition Count Per Year		3.75*	42.41	0.13	$<10^{-10}$	5.00*	148.51	0.44	$<10^{-10}$
Portion of Man-preferred Institution Exhibition		4.56*	4.79	0.11	$<10^{-10}$	-	-	-	-
Portion of Woman-preferred Institution Exhibition		0.01	1.01	0.15	0.92	-	-	-	-
Career Length		4.44*	85.61	0.13	$<10^{-10}$	4.91*	136.15	0.53	$<10^{-10}$
Gender (Baseline: Man)	Woman	-0.13*	0.88	0.04	0.001	-0.23	0.79	0.13	0.07
Co-exhibit. Gender (Baseline: Co-exhibit. Man)	Co-exhibit. Neutral	0.02	1.03	0.07	0.701	-0.71*	0.49	0.16	$<10^{-6}$
	Co-exhibit. Woman	-0.21*	0.81	0.07	0.003	-0.69*	0.50	0.14	$<10^{-8}$
Gender x Co-exhibition Gender (Baseline: Man, Co-exhibit. Man)	Woman, Co-exhibit. Neutral	-	-	-	-	-	-	-	-
	Woman, Co-exhibit. Man	-	-	-	-	-	-	-	-
	Woman, Co-exhibit. Woman	-	-	-	-	-	-	-	-
	Man, Co-exhibit. Neutral	-	-	-	-	-	-	-	-
	Man, Co-exhibit. Woman	-	-	-	-	-	-	-	-
Prestige (Baseline: High Prestige)	Low Prestige	-	-	-	-	-0.18	0.83	0.17	0.264
	Mid Prestige	-	-	-	-	-0.28*	0.75	0.14	0.046
Most Exhibited Continent (Baseline: North America)	Africa	-	-	-	-	-	-	-	-
	Asia	-	-	-	-	-	-	-	-
	Europe	-	-	-	-	-	-	-	-
	Oceania	-	-	-	-	-	-	-	-
	South America	-	-	-	-	-	-	-	-
Medium (Baseline: Painting)	Drawing & Watercolor	-	-	-	-	-0.34*	0.71	0.17	0.045
	Installation	-	-	-	-	-0.93*	0.39	0.15	$<10^{-9}$
	Mixed Media	-	-	-	-	-0.02	0.98	1.25	0.98
	Photography	-	-	-	-	-0.78*	0.46	0.21	0.0001
	Print	-	-	-	-	-0.49	0.61	0.67	0.46
	Sculpture	-	-	-	-	-1.17*	0.31	0.33	0.0004
	Video & Film	-	-	-	-	-1.92*	0.15	0.84	0.022

Other annoyances

While I have listed my major concerns about the paper above, there are a few other issues that I would recommend the authors address or at least consider as having caused at least one reader some concerns about the scientific quality of the research.

• The analysis focuses on about 65 thousand artists out of over 465 thousand – and a fully convincing explanation for constructing the sample is not provided. Why, for example, is consideration limited to those artists who began exhibiting on and after 1990. The explanation offered implies that this was required for accurate reconstruction of exhibition history, but I find this difficult to fully accept. The approach introduces potential sample selection bias into the analysis that should be acknowledged and discussed. Does this limit exclude Damien Hirst, for example (whose work featured in a group exhibition in 1988)? It surely would exclude Jeff Koons and many other "blue chip" contemporary artists, so some further explanation seems warranted.

To clarify, the primary objective of specifying a threshold for the start-year of artist careers was to exclude artists from antiquity that make up the bulk of 465 thousand artists originally mentioned in the dataset. It is our belief that the success of these artists is no longer influenced by the decisions at institutions, as many of these artists are renowned for their historical importance. Since the focus of this manuscript is the potential social influences on inequality and artist career success, we choose to limit the analysis to contemporary artists who are beginning their careers in the modern gallery/museum era, and for whom we have a greater chance of capturing full exhibition histories. Since there is no explicit definition for the year defining contemporary artists, we choose the year 1990. Importantly, in the SI Figure S2, we demonstrate that the rates of gender inequality do not change significantly if one chooses different years for the threshold, and all results remain qualitatively unchanged.

• The current text contains numerous problematic assertions about how exhibition venues respond to incentives to make their decisions. For example, on page 6 it is asserted that the quantitative analysis of gender representation could be offered to the gallery owners, museum curators and boards, then it would offer "… them the metrics and structural insights to empower them to offer balanced representation." I find it very difficult to believe that these decision makers lack the power change gender composition of the groups of artists they exhibit, nor that these quantitative measures will provide "metrics and structural insights" to these persons. It would be much better to think through a careful model of their decision making, use the metrics to test, refine, and validate the model, and then use the model to identify how to induce changes in their decisions.

We thank the Reviewer for their interesting suggestion of future directions to advance the policy analysis component of this work. Firstly, we note that we are not arguing that these decision makers lack the power to enact change. Rather, in our experiences talking with gallery and museum curators, we found there was a pressing need to provide data and tools for quantifying the extent of gender inequality in their exhibitions and collections. Surprisingly to us as well, many of these curators had not kept detailed records of historical exhibitions within their institutions, or when they did have records, those records did not contain the artist genders (an attitude that has greatly changed recently). Here, our argument is that they should be considering such data in their decision making processes. And we agree with the Reviewer that such numbers will have limited influence on galleries who serve collectors. However, such arguments and metrics do have increasing weight in public-focusing museums, even if they are private. Secondly, our work suggests that gender inequality in art is not only localized within individual institutions, but that there are facets which emerge from collective decisions at groups of institutions. These inequalities will likely require collective action to address. Finally, we wrote this discussion with the future research direction of building artistic recommendation tools that could expand the scope of artists an institution considers for exhibitions (controlling for relevant features such as prestige and medium). These tools require extensive data such as those we employ here. We now engage more explicitly with this idea in our discussion section.

> • On page 15 the text states that "Sales … and the resulting valuation of an artist's work, is a frequently used (and misused) measure of artistic success." This kind of "nudge and wink – do you get it?" argument appears frequently in journalism, but in my view has no role in scientific writing. "Misused" in what sense? Why is focus on the value of an artist's work a "misused" measure? Are artists supposed to not be paid? Is the assertion that for an artist to be "starving" is the natural order of things, or that for some reason artists are not supposed to be concerned about the valuation of their work, but only about the inner glow that they experience in its creation? The text doesn't say, but simply asserts that focus on valuation is frequently misused. If the authors have evidence of this, present it. If they have a clear argument to make or to test using their data, tell us.

We have now removed the collaqual reference from our presentation.

> • Another example of what might be characterized as "unsupported or arm-chair social theorizing" is presented in the penultimate paragraph of the paper, where the authors assert that the gender gap in exhibition of artworks is "can be only balanced out through the adoption of affirmative acquisition and perhaps gender affirmative programming – i.e. policies whereby institutions over-emphasize art by women artists in their exhibitions and purchases aiming to move them closer to gender balance over time." Do the authors have any evidence to support this claim? Any examples where such policies have been successful? I would suggest avoiding ending the paper with a claim that feels "progressive and relevant" but is unsupported by the evidence assembled or analysis presented. Such writing may be appropriate for Artnet but is probably best left out of serious scientific journals.

There is in fact recent evidence. For example, in their attempts to restore gender balance, The DC based National Museum of Women in the Arts is dedicated solely to women artists, the Baltimore Museum of Art is committed to only purchasing new works by women artists[1], the Netherlands based Rijksmuseum added women to its prestigious "gallery of honour"[2], and The

Frieze highlighted five new galleries which are dedicated only to women artists[3], to name a few. We have now softened the claim of affirmative acquisition as the only policy, since there may well be others we have not considered.

[1]https://www.washingtonpost.com/lifestyle/2019/11/15/baltimore-museum-art-will-only-acquire-works-by-women/

[2]https://www.smithsonianmag.com/smart-news/rijksmuseum-will-display-work-women-artists-its-gallery-honour-first-time-180977209/

[3]https://www.frieze.com/article/five-trailblazing-galleries-only-show-women-artists

• The current text states on page 23 that the data "… with artist gender and institutional measures of gender inequality [is] now available on the GitHub repository …." (emphasis added). As far as I can tell the GitHub repository does not contain the data, but only a 2-line readme file with no additional information.

We thank the Reviewer for catching this problem. We have now opened the public access on Github to the data.

References

Adams, R. B., Kräussl, R. Navone, M., & Verwijmeren, P. (2021). Gendered prices. The Review of Financial Studies, 34(8), 3789–3839.

Bloch, Francis & Dutta, Bhaskar (2009). Communication Networks with Endogenous Link Strength
Games and Economic Behavior, 66:39-56.

Bocart, F. Y., Gertsberg, M., & Pownall, R. A. (2021). An empirical analysis of price differences for male and female artists in the global art market. Journal of Cultural Economics, 1–23.

Brown, T. W. (2019). Why is work by female artists still valued less than work by male artists? www.
artsy.net/article/artsy-editorial-work-female-artists-valued-work-male-artists.

Cameron, L., Goetzmann, W. N., & Nozari, M. (2019). Art and gender: Market bias or selection bias? Journal of Cultural Economics, 43(2), 279–307.

Cowen, T. (1996). Why women succeed, and fail, in the arts. Journal of Cultural Economics, 20(2), 93–113.

Droege, J. (2022). First impression biases in the performing arts: Taste-based discrimination and the value of blind auditioning. Journal of Cultural Economics volume 46, pages 391–437.

Galenson, D. W. (2007). Who were the greatest women artists of the twentieth century? a quantitative investigation. NBER Working Paper, No. w12928.

Greenwald, D. S. (2021). Painting by numbers: Data-driven histories of nineteenth-century art. Princeton, NJ: Princeton University Press.

Hoffman, R., & Coate, B. (2020). Fame, what's your name? Quasi and statistical gender discrimination in an art valuation experiment. Working paper.

Huhn, Tom (1993). The Field of Cultural Production: Essays on Art and Literature. Journal of Aesthetics and Art Criticism January 1993.

LeBlanc, A. & Sheppard, S. (2022). Women artists: gender, ethnicity, origin and contemporary prices. Journal of Cultural Economics, 46:439–481.

Olaizola, Norma & Valenciano, Federico (2020). Characterization of Efficient Networks in a Generalized Connections Model with Endogenous Link Strength, SERIEs, ,11:341-67, September 2020.

In summary, we thank the Reviewer for their detailed comments and recommendations, all of which have greatly helped to enhance the clarity of our manuscript.

Reviewer #1

Thank you for your detailed responses to my comments. I appreciate the additional analyses.

However, I believe there is an opportunity for further discussion about what galleries and museums might do to address the gender gap. While I understand that your aim is not to prescribe specific actions for these institutions, your study clearly highlights a worrying pattern that needs addressing.

This addition would enhance the practical relevance of your research and might serve to guide those in the art world. I would suggest you reconsider adding a section discussing possible approaches that institutions could adopt or explore further to address the gender gap you have highlighted in your study. To that end, I recommend considering the findings and recommendations of the following papers, which, while not directly related to the art world, offer valuable insights into successful strategies for reducing gender biases and promoting equality:

- Carnes, M., Devine, P. G., Manwell, L. B., Byars-Winston, A., Fine, E., Ford, C. E., ... & Sheridan, J. (2015). Effect of an intervention to break the gender bias habit for faculty at one institution: a cluster randomized, controlled trial. Academic Medicine: Journal of the Association of American Medical Colleges, 90(2), 221.
- Kossek, E. E., Su, R., & Wu, L. (2017). “Opting out” or “pushed out”? Integrating perspectives on women’s career equality for gender inclusion and interventions. Journal of Management, 43(1), 228-254.
- Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., ... & Greenwald, A. G. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. Proceedings of the National Academy of Sciences

We thank the Reviewer for the suggestion and we have followed it by adding a discussion of possible approaches to address the gender inequalities in the art world that we identified in our work.

“Our work opens new directions for future research into potential interventions that may address the gender inequalities in the art world. For example, in education and academe, evidence-based training has been shown to reduce individual stereotyping and implicit bias (Carnes et al., 2015; Kossek et al., 2017). Due to the documented network constraints in artistic careers, these interventions may have propagating influences beyond target institutions which would enhance their effectiveness. Our findings on gender preference lock-in also suggest that potential interventions offering greater resources or opportunities to exhibit early in an artist’s career may have compounding influences.”

Reviewer #2

> I have taken a look at the revision and at the very detailed responses of the authors. I appreciate the careful attention to the feedback and I am convinced by the revisions. I'd be very happy to see this important paper published.

We would like to express my sincere gratitude to the Reviewer for their positive comments and for their valuable feedback on previous submissions.

Reviewer #3

> The authors have adequately addressed my comments. I do not have any further major comments.

We are pleased to hear that our revisions have effectively addressed your previous comments.

Reviewer #4

I appreciate the authors' efforts to address some of the comments.

I have the following additional comments/concerns.

(1)With the additional references added to the revised manuscript, I am not convinced that the findings of this study are novel enough. There have been similar studies with similar findings for the art world. The major difference between this main study and others might be that this study used a larger dataset than many of them.

We thank the Reviewer for providing another opportunity for us to situate our work within the burgeoning and multifaceted fields of gender inequality and quantitative art studies. Our project leverages both distinctive data and methodologies which empower novel analysis to reveal the systemic forces that perpetuate gender inequality in art and quantify the influences on artist careers. Specifically, our data contains the most detailed exhibition information from both museums and galleries used so far in analysis, full artist exhibition careers, artist demographics, and their sales at auction. The references we added, namely Adams et al. (2021), LeBlanc & Sheppard (2022), and Bocart et al. (2022), focus only on auction sales, artist demographics, and artwork characteristics, but have no access to exhibition data and thus are unable to draw insights into how prestige and network effects influence gender inequalities in art. Critically, none of these studies focus on art institutions or exhibitions, a key feature of our work, and hence are unable to reveal how art institutions influence artist careers by limiting long-term access to resources and rewards in art. Since there is nuanced and complex interplay between artistic production, museum exhibition, gallery exhibition, and auction sales, our study constitutes a new and definite direction in the study of gender inequality in art.

(2) For the data, I'm concerned about the selection of artists included in the analytical sample. The manuscript indicated that it identified the gender category of 471,791 artists, but only kept 86,894 (18%) of them after removing individuals who started their careers after age 50 or before age 18. The study was then further limited to those who started their exhibitions on and after 1990. It was not reported in the study regarding the accuracy of the demographic information. It could be possible that those excluded exhibit certain patterns by gender. The further exclusion of artists who started on or after 1990 might suffer from incomplete coverage of individual careers, and their pattern based on gender (if any).

We share the Reviewer's concern for clarity in sample preparation which originally motivated us to clearly underscore the artist population in our project's title. It's important to note that the original dataset covers exhibitions of artwork from all periods of art, including those containing artwork from antiquity (e.g. Greek, Roman, Egyptian, etc.), Renaissance, Cubism, etc.. Hence, as originally argued in our SI, the majority of artists in our dataset are likely more renowned for their historical importance than for the social influences of institutional decisions. Since the focus of this manuscript is the potential social influences on inequality and artist career success, we choose to limit the analysis to contemporary artists who began their careers in the modern gallery/museum era. We choose the year 1990, often argued as the beginning of the modern gallery system (Szanto, PhD Thesis 1996). Importantly, in the

SI Figure S2, we demonstrate that the rates of gender inequality do not change significantly if one chooses different years for the threshold, and all results remain qualitatively unchanged. Similarly, prompted by this Reviewer's concern, we have now added SI Figure S8 which includes another robustness check in which we remove the age criteria but retain the first exhibition date criteria. All of our original results remain qualitatively similar.

Since the demographic information is expert curated, we do not have an additional means to validate the artist's age at scale. However, we randomly selected 20 artists and verified the birthdate matched their wikipedia biographies for all 14 cases which had a public biography. We could not find a public resource to validate or invalidate the remaining 6 cases.

> (3) I'd suggest avoiding using words to imply causality in the manuscript. For instance, on Page 3: "This focus is motivated by our goal of identifying systemic influences of gender representation on active artist careers;". There are many other similar instances throughout the manuscript.

We agree and we have now revised the manuscript to address our language.

> (4) For the disparities in exhibition opportunity, the manuscript indicates the gender ratio is 1.86, the larger the gender ratio in the whole sample (1.74). Yet the mean difference between the exhibition opportunities of men and women is 0.9 (6% difference). Because productivity is not accounted for here, it is impossible to estimate how much of these disparities are due to productivity. It is also not clear here how sensitive the conclusion is to certain individual artists being extreme cases (i.e., certain super popular artists with many exhibitions).

We thank the Reviewer for their concern over exhibition disparities, and agree that we do not have enough evidence to draw insights into its origin. Instead, our work focuses on the constraints imposed by social structure underlying the art world and their consequences for institutions and artist careers. For this analysis, we control for the number of exhibitions to remove its potential influences.

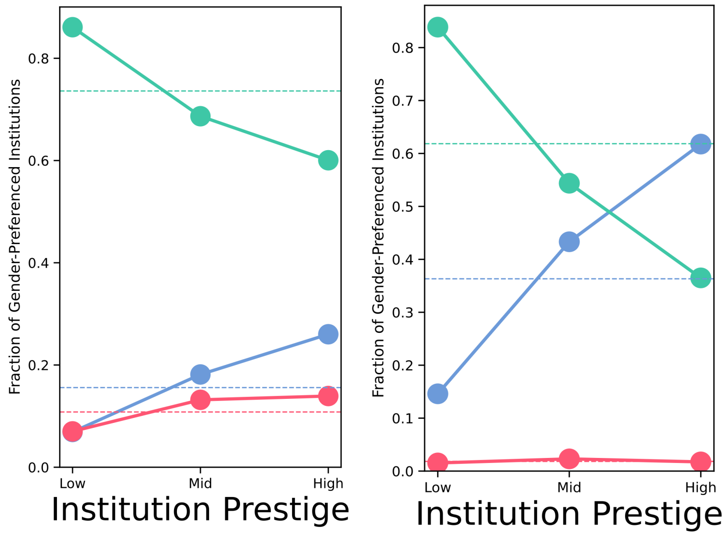
> (5) Maybe the concept of gender-balance and gender-neutrality should be introduced early on.

We introduced the concepts when they become essential for the analysis after finding that starting the full section with these definitions just confused the reader.

> (6) In the analyses based on institutions, the manuscript indicated that 62.14% of institutions were excluded due to the limited number of exhibitions (n<30), which is about 10% of the total. What's the gender composition of those exhibitions? If larger/more prestigious institutions are more likely to favor men, whether/how does the excluding of these exhibitions speak to the conclusions here?

To clarify, we retained all institutions for calculating institutional prestige, and only excluded the smallest institutions (by exhibition count) for the classification of institutional gender preference. Motivated by the Reviewer's concerns, we ran an additional robustness check in which we removed the previous exhibition count n>=30 criteria. Under the gender-neutral criteria, this results in more gender-neutral institutions since the vast majority of small institutions do not have statistical significance. In the full dataset, we find 15,008 gender-neutral institutions, 3,177 men-preferred institutions, and 2,204 women-preferred institutions. Most importantly, as we show below and in the SI, Figure S6 d, h, we retain the observation of more men-preferred institutions at high-er levels of prestige.

Gender-preference vs prestige (left: gender neutral, right: gender balance)



(7) For the regression analyses on auction opportunities, I find the statement “Overall, we find that the gender preference of institutions embracing an artist’s work is a stronger indicator of the likelihood that an artist will transition to the auction market than the gender of the artist, capturing an institutionally driven gender premium for men” a bit disturbing. I don’t think coefficients of different independent variables/models can be directly compared.

We now rephrased this sentence to highlight that the conclusion is based on the odds ratio not the coefficients, and also note that both variables are binary with coefficient differences outside of the standard errors.

(8)The references seem to be messed up: the first reference is numbered 29. I thought the references appear in number order?

We have corrected the reference numbering.

References

Szanto A. *Gallery transformations in the new york art world in the 1980's.* [Order No. 9631787]. Columbia University; 1996.

I was invited to focus on the Bayes factor and sensitivity analysis part of the manuscript. After reviewing the manuscript and the associated GitHub scripts regarding the Bayesian Factor (BF) analyses, I have two concerns requiring clarification:

1. Prior Justification
The authors selected a beta(1,1) prior for their analysis but did not justify this choice. Given the critical role of priors in Bayesian inference, a rationale for selecting this uniform prior—or a sensitivity analysis comparing alternative priors (e.g., beta(0.5,0.5) or weakly informative priors)—should be included to support the robustness of the conclusions.

> We appreciate the Reviewer's attention to the role of priors in Bayesian inference. We originally selected a Beta(1,1) prior as a non-informative (uniform) baseline to reflect a lack of prior knowledge and to allow the data to primarily drive the inference. However, we agree that transparency and robustness are important. In response, we have conducted a sensitivity analysis using alternative priors, including Beta(0.5,0.5) and Beta(2,2) to examine the robustness of our classifications. These new results are now presented in the Supplementary Information, Section S2.5 and Figure S5. We found that the classifications remained fairly consistent across these priors, indicating that our conclusions are not sensitive to the specific prior chosen.

2. Classification Sensitivity and Simulation Transparency
The authors categorized institutions using a BF10 threshold of 1/3 and stated in their response that "with five exhibition observations, institutions are correctly classified as gender-neutral 85% of the time." However, there were two issues:
2.1 Documentation Gaps: The GitHub code does not explicitly include the simulation script to generate this 85% accuracy claim. To validate this, I attempted to replicate a simplified scenario:
• Ground truth: 50% male artists (p = 0.5), n = 5 exhibits.
• Simulation result: 0 of 1000 trials met the BF10 < 1/3 threshold for correct classification (true negative); and 4 of 1000 trials met the BF10 >= 3 threshold (false positive)

> We thank the Reviewer for their close reading and detailed review of the code. The simulation was developed during the course of the review process and had not yet been incorporated in the GitHub repository at the time of your inspection. We have now updated the repository to include the simulation. The link to the note book is here.

2.2 Effect Size and "hit rate": The above simulation used the parameters from the model of H0. It is also important to simulate the scenario with H1. I tried two effect sizes (p = 0.8 and p = 0.55), with n = 5, the percentage of trials that met the BF10 > 3 threshold for the correct classification was 31% and 6.2%, respectively. Overall, my simulation suggests n = 5 is too small to detect either the true negative or true positive events.

To ensure reproducibility and validate the claimed performance, the authors should:

- Share the complete simulation code used for accuracy calculations.
- Report sensitivity analyses across varying n (number of exhibits) and effect sizes (i.e., the deviation of p from 0.5 or 0.364).

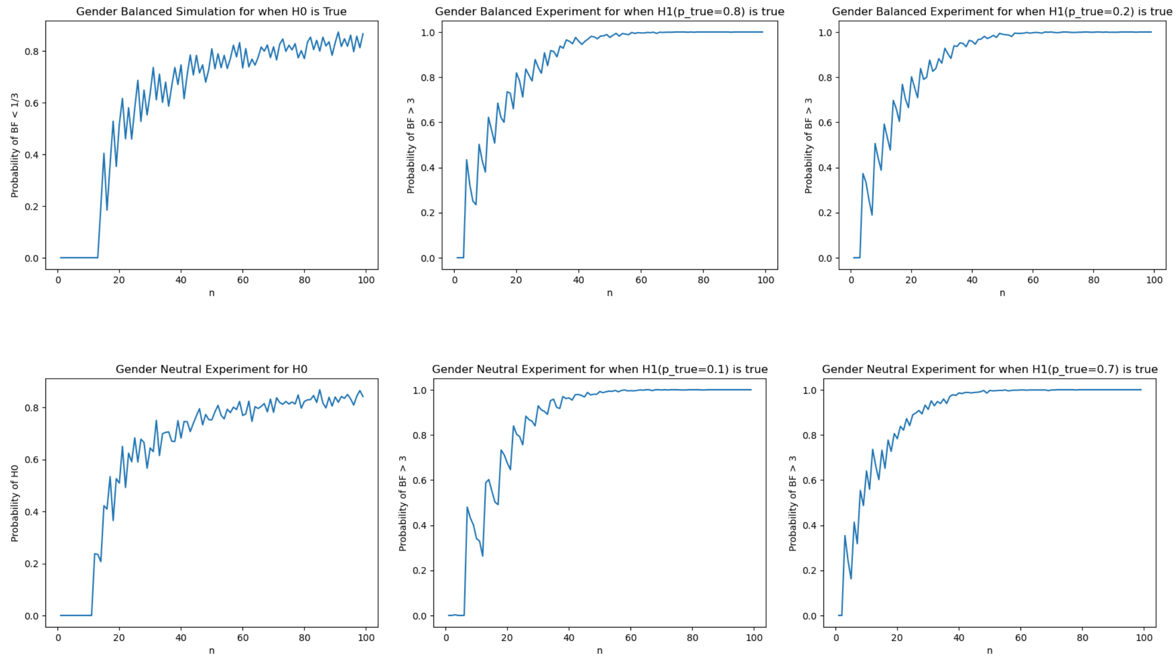
Below is my simulation code for reference:

```
# the code was based on JASP team's R code, the results are similar to the authors' (slightly
different precision)
def bayes_factor(counts, n, a=1, b=1, theta0=0.5):
logBF10 = betaln(counts + a, n - counts + b) - betaln(a, b) - counts * np.log(theta0) - (n - counts)
* np.log(1 - theta0)
BF10 = np.exp(logBF10)
return BF10

n = 5
p = 0.5 # also simulated with 0.365, 0.55, 0.8
x = np.random.binomial(n, p, 1000)

y = bayes_factor(x, n)
proportion = np.sum(y < 1/3) / 1000
print("Proportion of Bayes factors < 1/3:", proportion)
```

We sincerely appreciate the Reviewer's thorough engagement with the simulation logic and the efforts to independently replicate our analysis. The comments have been extremely helpful in improving the transparency and rigor of our work.

Regarding the concerns raised about the simulation under the alternative hypothesis (H1), we uncovered a minor error in our original specification of H1 that affected the calculation of "hit rates." This led to overly optimistic accuracy estimates in some scenarios. We have now corrected this issue and updated the simulation code accordingly. The revised simulations, along with documentation, have been added to the GitHub repository as of this revision, contained in the jupyter notebook: BF_accuracy_simulation.ipynb. We now conduct the simulation for both H0 and H1 across the two evaluation criteria and report the results of the sensitivity analysis in the SI, Figure S6, and reproduced below.

While classification accuracy is somewhat lower for smaller institutions, this does not substantially impact the overall reliability of the model. The majority of artists exhibit in larger institutions, which dominate the classification signal and ensure consistent categorization. To further assess robustness, we conducted an additional classification experiment using a frequentist approach with an n>30 threshold. The resulting models were qualitatively consistent with our Bayesian results, reinforcing the stability of our findings across methodological frameworks.



**

After attempting to reproduce the full code and re-read the methods and results sections, I have two comments for the frequentists statistics in the results section:

(1) On page 8, the comparison between museums and galleries is only presented as descriptive data; no statistical test was presented here. I am not sure the authors can claim one is higher than the other, e.g., "3.76% of galleries vs 2.55% of museums" does not necessarily demonstrate a "higher fraction of galleries ... than museums".

We thank the Reviewer for raising this important point. In the original manuscript, we reported the observed proportions descriptively (3.76% for galleries vs. 2.55% for museums), based on our sample. We agree that without a formal statistical test, such a comparison could be misinterpreted as implying a definitive difference.

To address this, we have now reframed the comparison as a Bayesian hypothesis test for the difference between two proportions. Specifically, we computed the Bayes factor comparing a null hypothesis H_0: The proportion is equal for galleries and museums and the alternative hypothesis H_1: The proportions differ between the two groups. Using a default Bayesian approach with a Beta(1,1) prior on the proportions, we find a Bayes factor of BF_{10} = 2.82, which provides weak to moderate evidence in favor of H_1. Full details of this analysis are provided in the updated Methods section and the corresponding code is now included in the GitHub repository.

This reframing allows us to quantify the uncertainty around the observed difference and avoid over-interpreting what may be due to sampling variation. We have revised the text accordingly to reflect this statistical framing.

(2) In the series of logistic regressions, the model comparison index used was BIC, but no numeric results were presented showing the magnitude of BIC reduction.

We appreciate the Reviewer's comment regarding the presentation of Bayesian Information Criterion (BIC) values. To clarify, we report the BIC values for each model in Table 2 of the main text and in Supplementary Information Table 3. These tables provide the numeric BIC values that were used for model comparison, allowing readers to assess the relative fit and complexity of each specification.

Specifically, we note that the BIC decreases from 20415.93 in Model 1 to 19755.10 in Model 3, a reduction of over 660 points. This represents very strong evidence in favor of Model 3 under standard BIC interpretation guidelines. To improve clarity, we have updated the text to highlight this magnitude of improvement and guide the reader to the relevant tables.

As the data is available now, I was able to run most of the script, but I tried for a while to get the files in the right folder and create new folders with the specified names, yet I could not run all the code. Here are some suggestions regarding the provided script and data: first, I recommend the authors share the data along with the code, which will make it easier for future readers to reproduce the analysis—I personally do not believe there is a huge difference between sharing via DataVerse and GitHub; second, I recommend the authors add info regarding the computational environment used for the analysis (i.e., version of Python and Python modules, operating system); third, more comments are needed to navigate readers, especially explaining the different pipelines; finally, while some prefer Python scripts for speed, Jupyter Notebooks are more reader-friendly for understanding the data analysis flow when the data volume isn't extremely large.

We thank the Reviewer for the effort to reproduce our results. We fully agree that accessing the data via DataVerse may be less seamless than alternatives, but this was sadly not our choice: arrangement was made in accordance with the requirements of our data provider, and we are not permitted to redistribute the data through other channels.

Regarding the suggestion to use Jupyter Notebooks, we appreciate the recommendation. However, because our analysis includes a large number of robustness checks with varying configurations, we chose to structure the code as Python scripts to reduce duplication and improve modularity. That said, we recognize the importance of clarity and reproducibility. In response, we have added additional comments and guidance within the scripts to help users navigate and execute the code more easily.

In general, I recommend not using two systems of hypothesis testing to avoid contradiction (see references below). In this manuscript, the Bayesian and frequentist statistics are used for testing different hypotheses, which I find acceptable. For the part where statistical testing was missing ('3.76% of galleries vs 2.55% of museums'), I also recommend using BF. If the authors are interested in using Bayesian analysis for the series of logistic regressions, they may consider using Bambi, a powerful Python module for Bayesian statistics.

Dienes, Z. (2024). Use one system for all results to avoid contradiction: Advice for using significance tests, equivalence tests, and Bayes factors. Journal of Experimental Psychology: Human Perception and Performance, 50(5), 531–534. https://doi.org/10.1037/xhp0001202
Schreiner, M. R., & Kunde, W. (2025). A cautionary note against selective applications of the Bayes factor. Journal of Experimental Psychology: General, 154(1), 294–304. https://doi.org/10.1037/xge0001666

We thank the Reviewer for the comments and shared resources. As suggested, we now used BF for the gallery and museum comparison.

Reviewer #6 (Remarks on code availability):

I reviewed the code for the beta-binomial model for BF (https://github.com/Barabasi-Lab/artGender/blob/main/codes/03-03-gender_preference_scatter_boundary.py), it seems to me that only partial script was shared, because the results of BF were loaded from json files ("gender_neutral_country_bf10") that were not shared on github.

We thank the Reviewer for the detailed review of our code. Regarding the specific concern raised, the file `gender_neutral_country_bf10.json` is generated by running the preceding script `03-00-gedner_preference_assign.py`. To ensure the codebase is fully reproducible, we have carefully reviewed all shared scripts and updated the readme.md with clearer instructions for running the pipeline, including the generation of intermediate files.

In summary, we are grateful for the Reviewer's close attention to the statistical and computational aspects of our work. Their detailed feedback led to improvements in the clarity of our statistical framing, the completeness of our simulation analyses, and the transparency of our codebase. These contributions have significantly strengthened the rigor and reproducibility of the project.