Preliminary and Incomplete – Do Not Post, Cite, or Circulate

Patient Identity and Support for Medical Care:
Evidence from List Experiments on Puberty Blockers*

Billur Aksoy♥, Christopher S. Carpenter♦, Danya Lagos♠, Dario Sansone♣

September 2026

Abstract

We study whether people support the same medical treatment differently depending on *who* receives it and *why* they receive it. Using double list experiments in the US and the UK, we measure views about puberty blockers for: 1) kids with precocious puberty; 2) transgender kids with precocious puberty; and 3) transgender kids with gender dysphoria. Support is lower when kids with the same condition – precocious puberty – are transgender. There is no difference in support across very different medical conditions for transgender kids. Public opinion thus reflects judgments about identities more than medical diagnoses. We also find significant social desirability bias.

Keywords: identity, discrimination, medical care, list experiment

JEL: C9, I1

* We thank seminar participants at the AEA CSQIEP Virtual Seminar, University of Exeter, and Vanderbilt University for their feedback. This project received IRB approval from Rensselaer Polytechnic Institute, University of Exeter, and Vanderbilt University (Study #00000046). The study was pre-registered with the AEA (AEARCTR-0016666). Financial support from the British Academy/Leverhulme Small Research Grants (SRG2223\231221) and Vanderbilt University is gratefully acknowledged. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY NC ND) license to any Author Accepted Manuscript version arising. The views expressed in this paper are those of the authors and should not be attributed to any funders or organization. All errors are our own.

♥ Rensselaer Polytechnic Institute. E-mail: aksoyb3@rpi.edu

♦ Vanderbilt University and NBER. E-mail: christopher.s.carpenter@vanderbilt.edu (corresponding author)

♠ UC Berkeley. E-mail: dlagos@berkeley.edu

♣ University of Exeter and IZA. E-mail: d.sansone@exeter.ac.uk

1. Introduction

Do people support the *same* medical treatment differently depending on *who* receives and *why* they receive it? Economists have long documented differential treatment of individuals based on their social identities. Yet, less is known about how these identities shape attitudes and support over the legality of receiving medical care. We aim to understand the roles of *patient identity* and *medical condition* in shaping public support in the United States and the United Kingdom for a specific type of medical treatment: puberty blockers. Puberty blockers are pharmacological treatments for a wide range of conditions affecting kids independent of their gender identity. These drugs temporarily stop pubertal growth and are reversible: once a patient stops taking puberty blockers, normal pubertal development continues. Puberty blockers are commonly prescribed for a condition called 'precocious puberty' which is a medical condition that can be experienced by any child and occurs when a child experiences unusually early pubertal development. Precocious puberty causes a mismatch between physical development and emotional and cognitive maturity and it has been linked to, for example, future risk of various mental health disorders (e.g., Hueg et al, 2026). Puberty blocker therapy is currently recommended as the primary pharmacological treatment for many children with this medical condition. Using a double list experiment method, we aim to understand whether public support for this medical treatment depends on *patient identity,* specifically when we ask about 'kids' (experiment 1) or 'transgender kids' (experiment 2). This design allows us to study whether a medically identical treatment–same drug, similar reversibility– under the same conditions is perceived to be legitimate or illegitimate purely based on the identity of the patient. Puberty blockers are also commonly prescribed off-label for the treatment of gender dysphoria, which is a mental health condition involving emotional distress caused when one's gender does not match their sex at birth. Gender dysphoria is the most common diagnosis among people who identify as transgender, including children. Correspondingly we also aim to understand whether support for the same medical treatment depends on the *medical condition* by asking about support for puberty blockers for transgender kids with unusually early puberty (experiment 2) versus transgender kids with gender dysphoria (experiment 3).

Transgender healthcare for youths - and transgender rights more generally - is one of the most controversial social issues in both the United States and the United Kingdom. Although most estimates indicate that about one percent of the population is transgender, a disproportionate amount of popular debate and policy discussion focuses on transgender people. As one example, the United States Supreme Court issued four separate decisions related to transgender youth over the past two years (relating to gender affirming care, sports participation, education, and 'conversion therapy'), with each of those four decisions seen as victories for advocates of a more restrictive approach to transgender rights. The United Kingdom has also been home to much debate about transgender youth and access to healthcare, as the UK National Health Service commissioned a highly influential report called the Cass Review which concluded that the evidence base for pediatric gender care for transgender youth was 'remarkably weak' (Cass 2024). The Cass Review played prominently in recent decisions by the UK and other European countries

to restrict access to puberty blockers for transgender youth, and the study was cited extensively in the 2025 US Supreme Court decision in *US vs. Skrmetti* which held that Tennessee's law prohibiting gender affirming care for minors did not violate the Constitution's Equal Protection Clause.

Given the magnitude of the public and political debate surrounding gender affirming care for transgender youth, understanding public attitudes towards these policies and issues is important. However, measuring public opinion on such a politically and socially sensitive topic presents a methodological challenge. In particular, social desirability bias may lead respondents to misreport their true views by providing answers that they deem to be more socially acceptable. Thus, conventional survey questions may not accurately capture respondent's true views. We implement a double list experiment technique, which is a commonly used method that allows researchers to elicit true views in the absence of social desirability bias (Miller 1984; Raghavarao and Federer 1979; Stantcheva 2022). We recruit several thousand participants from Prolific who are representative of the United States and the United Kingdom across race, sex, and political affiliation.

We report several key findings. First, we show that for a given type of medical care, public support is impacted by patient identities. Specifically, we show that true support for puberty blockers for unusually early puberty is significantly lower when we ask about 'transgender kids' versus 'kids'. The magnitude of the decline in support for transgender kids is similar across the US (10.4 percentage points) and the UK (11.6 percentage points). Second, we show that in the two list experiments where we ask about transgender kids, true public support for access to puberty blockers does not vary across two very different medical conditions: precocious puberty and gender dysphoria: we find only 1.2 and 1.5 percentage point differences across these two conditions in the US and the UK, respectively, and neither estimate is statistically significant.

What explains these patterns? We argue for a mechanism we term *criterion-category conflation*. Through this mechanism, we suggest that if a specific diagnostic criterion becomes deeply associated with a contested social category, true public support is driven by the identity category rather than the stated content behind the criterion. In our context, participants' true support for the treatment of precocious puberty falls by roughly ten percentage points when the child is described as transgender in comparison to a child being treated for precocious puberty who is not described as transgender. On the other hand, once a child is identified as transgender, there are no meaningful differences in true support for the legality of prescribing puberty blockers to a transgender child living with precocious puberty and a transgender child living with gender dysphoria. Criteria for treatment become irrelevant in the face of a social category once it is named.

We also find that there is significant social desirability bias across all three experiments, ranging from 9.5 to 17.1 percentage points in the US and 8.6 to 31.4 percentage points in the UK. This suggests that common public opinion polls that ask questions about the use of puberty blockers

and gender affirming care for minors are likely to significantly overstate the true level of support. Interestingly, when we compare true views elicited from the list experiment with views from the direct questions about our key statements of interest, we find that people's private views are much more dependent on patient identities as compared to medical conditions than are their public views. In fact, in the UK – which has experienced public debate about gender affirming care for transgender minors for a longer period than in the US and which is home to the Cass Review that has influenced National Health Service policy on this care – we find no meaningful difference in stated support of prescribing puberty blockers for precocious puberty across patient identities but large differences in privately held true views across those same conditions. Conversely, the UK sample exhibits a large difference in stated support for prescribing puberty blockers based on the criterion applied, but our experiment finds that their privately held true views track the category of the patient, rather than the criteria being applied.

We also provide evidence on people's beliefs about population support for each of our three key statements, showing that people's beliefs are close to true population levels of support as implied by the list experiments. We provide descriptive evidence on the reasons participants list as the strongest factors explaining lack of support for access to puberty blockers for transgender youth, finding that concerns about regret and uncertainty about the long-term consequences of such care are the most prominent as reported by participants in the UK while uncertainty about the long-term consequences of such care and beliefs that political ideology are driving these decisions are the most prominent as reported by participants in the US. Religious concerns are much more likely to be voiced as a key determinant of views in the US compared to the UK, which is a much more secular country. We also compare views about gender affirming care for transgender youths to those for a wide range of other transgender rights issues. In both settings we find that stated support for access to puberty blockers for transgender youth is about the same as stated support for the ability of transgender people to change their gender on government documents but is lower than stated support for transgender people in the military and support for bans on so-called 'conversion therapy' for LGBTQ youth. In both settings we also find that stated support for access to puberty blockers for transgender youth is higher than the share of participants who state that parents should not be able to opt-out their children from instruction with LGBTQ+ books and the share of participants who state that parents should not have to be notified by teachers if children ask to use pronouns that are different from their sex at birth.

Finally, we demonstrate important differences across political ideology in stated and true views about gender affirming care for transgender youth. Specifically, we show that progressive participants (i.e., Democrats in the US and Labour party in the UK) are much more likely to support access to puberty blockers for transgender youths than conservative participants (i.e., Republicans in the US and the Conservative party in the UK), which is not surprising. What is perhaps more surprising is that the magnitude of social desirability bias among progressive participants is much larger - on the order of twice the magnitude - than the SDB exhibited by conservative participants. Beliefs also exhibit interesting political heterogeneity: conservative participants underestimate the

true level of population support for access to puberty blockers for transgender youths, while progressive participants overestimate the true level of support. In contrast, median beliefs of Independents in the US are almost exactly identical to the true level of population support in the country.

Our study extends research on identity-sensitive policy attitudes - or support for policies that are driven by sentiments towards the identity groups that are perceived to benefit from them (Stantcheva 2026; Brown 2013; Cottrell, Richards, and Nichols 2010; Fox 2004; Nelson 1999; Wetts and Willer 2018). We address this phenomenon beyond its traditional basis in the relationship between race and welfare politics into the increasingly contested domains of health care and gender identity. Our results are also important because major policy decisions in both the US (the Tennessee law at issue in *US vs. Skrmetti*) and the UK make explicit carveouts for precocious puberty and thus presume to be implementing an objective, category-neutral barometer based on medical indication corresponding to diagnostic criteria. However, our findings suggest that the public's willingness to permit treatments largely flows from the type of person who is imagined to receive it.

Our work builds on several key literatures in economics. First, we contribute to the broad economics literature on identity and discrimination. Within this literature, our paper specifically contributes to a health economics literature on the role of identity in healthcare delivery and receipt. Many studies examine the role of racial and ethnic identity and concordance in healthcare, finding important differences for racial minority patients (e.g., Alsan et al. 2019, Alsan et al. 2026, Angerer et al. 2019, Fumarco et al. 2024, Obermeyer et al 2019, Singh and Venkataramani 2026). Few studies also considered gender, specifically the transgender identity, of health care recipients (e.g., Fumarco et al. 2024). Our paper expands this literature by focusing on public support for the same medical treatment across *who* receives and *why* they receive it.

Second, we add to a growing body of evidence in economics on transgender, nonbinary, and other gender minority people (see Badgett et al. 2024 and Aksoy et al. forthcoming for reviews). Most prior work has focused on earnings and income for gender minority adults compared to similarly situated cisgender people and has found very large gender minority income penalties (Carpenter et al. 2024, Carpenter et al. 2023, etc.). Some experimental economic research examined the economic traits (such as confidence, competitiveness, and risk tolerance) of gender minority individuals (including gender minority youth) compared to cisgender people (Aksoy et al. 2024 and Coffman et al. 2026). Other work has examined the roles played by policy in shaping adult health and socioeconomic outcomes, such as laws permitting gender marker changes on government documents (Mann 2021) or Medicaid coverage for gender affirming care (Mann et al. 2024). Fewer studies have examined transgender youth, mainly due to data challenges, though prior work has used siblings designs to examine the role of family background in producing mental health disadvantage for gender minority adolescents (Carpenter et al. 2026). Regarding use of puberty blockers and other gender affirming care by youths, most of this work is published in

medical and public health journals, though a small number of studies leverage plausibly exogenous variation in the timing of initiation of gender affirming care to examine arguably causal effects on mental health and labor market outcomes (Saxby et al. 2026, Campbell 2025, Campbell et al. 2026, and others). We are not aware of any prior work in economics that examines the levels or determinants of public support for puberty blockers or other gender affirming care for minors.

Finally, our study is related to a sizable literature in economics that uses list experiments to understand stigmatized behaviors and identities, including LGBTQ+ identities. In the context of LGBTQ+ identities and topics, Coffman et al. (2016) use a list experiment to show that the self-reported LGBT identity is significantly underreported, and they also show that stated support for same-sex marriage and other LGBTQ-related policies is overreported. Aksoy et al. (2024) study comfort with transgender managers and support for transgender-inclusive nondiscrimination policy; using a double list experiment approach, they find evidence of social desirability bias despite a strong majority of participants expressing true comfort with having a transgender manager and support for nondiscrimination. Listo et al. (2026) study comfort with gay people in Chile as reported by supervisors, coworkers, and customers, finding evidence of significant social desirability bias, especially among supervisors.[1]

2. Experimental Design

In this section, we discuss our experimental design and protocol. Our pre-registration and pre-analysis plan can be found on the American Economic Association's registry for randomized control trials (AEARCTR-0016666): https://doi.org/10.1257/rct.16666. The complete set of instructions and survey questions used for our study can be found in the Online Appendix.

The main analysis relies on a double list experiment technique (also called "item-count technique", "unmatched count", or "veiled approach") that was pioneered by (Raghavarao and Federer 1979) and (Miller 1984). In a regular list experiment, participants are either assigned to a control or a treatment group. In the control group ("short list"), participants are given a list of statements and asked to indicate how many of those statements are true for them. In the treatment group ("long list"), participants are given a list which includes the same set of statements as the control group plus one additional statement (i.e. the key statement). Since the only difference between the control and treatment group is the key statement, the difference in means in responses between the treatment and control group is the estimated share of the sample with that key sensitive attribute.

In our case, we exploit the double list experiment technique to increase power, following Glynn (2013). For a given key statement, we use two lists (i.e., List A and List B). Some participants are

[1] There are numerous examples outside of LGBTQ+ topics, including in adjacent disciplines; for reviews and meta-analyses, see Li and Van den Noortgate (2022) and Blair et al. (2020). Within economics, Alsan et al. (2026) study the demand for lethal firearms and use list experiments to uncover participants' true reasons for owning firearms, while Bursztyn et al. (2020) use list experiments to understand women working outside the home.

randomized into seeing List A (a short list) and List B + Key Sensitive Statement (a long list). Other participants are randomized into seeing List A + Key Sensitive Statement (a long list) and List B (a short list). We calculate the differences-in-means between the short and long lists for both A and B lists, then take the average. This average of the differences-in-means is the share of our sample with that key sensitive attribute. A more thorough discussion of the double list experiment technique can be found in Aksoy et al. (2025). Table 1 presents the non-key as well as key sensitive statements used in the experiment.

Given that we use a double list experiment, for each key sensitive statement there are four possible list orders. The first list order is that a participant sees List A (short list), then List B with the Key Sensitive Statement (long list). The second list order is that a participant sees List A with the Key Sensitive Statement (long list), then List B (short list). The other two list orders are those with the order of these lists reversed. Since we have a total of three key sensitive statements, this results in a total of 12 list orders. The complete set of list orders can be seen in Online Appendix A. We employ full randomization: participants are randomized into one of these list orders and thus each participant encounters one and only one of the three key sensitive statements. As described above, Key Sensitive Statement 1 asks participants whether they think it should be legal for kids with unusually early puberty to obtain puberty blockers to treat that condition with parental consent. Key Sensitive Statement 2 is identical to Key Sensitive Statement 1 with the exception that we refer to 'transgender kids' instead of 'kids'. Key Sensitive Statement 3 is identical to Key Sensitive Statement 2 with the exception that we replace 'a medical condition that causes unusually early puberty' with 'gender-related psychological distress'.

After the list experiment, participants complete a survey where we collect standard demographic and socio-economic variables as well as opinions on various LGBTQ+ related issues. We provide more details on these questions as we discuss them later in the results section. Importantly, we ask the key direct question corresponding to the key sensitive statement each participant encountered as part of the list experiment. By comparing the average support obtained from the list experiment to the average response to the direct question, we estimate the size of the social desirability bias for each key sensitive statement.[2]

2.1. Key Design Considerations

In designing and implementing the list experiment, we carefully followed the best practices in the literature (Glynn 2013, Aksoy et al. 2025). Following Aksoy et al. (2025), we included four non-key statements as follows: one of the statements was expected to be true for most people to avoid

[2] Before the launch of the main data collection, we collected pilot data to test the statements and the experiment in general and to obtain additional insights. Following our pre-registration, participants were randomized into either participating in the full study described here or only seeing the survey. We find evidence that seeing the list experiment significantly impacted the responses to the survey in the UK. As a result, when we analyze the second list experiment in the UK we only use direct survey responses for those who did not see the list experiment. For the rest of the statements we do not find any significant evidence of an order effect and thus our study sample includes all participants.

a floor effect (i.e., participants reporting zero items which removes the privacy protection provided by the list experiment), one of the statements was expected to be false for most people to avoid a ceiling effect (i.e., participants reporting that all statements are true for them which removes the privacy protection provided by the list experiment), and the remaining two non-key statements are expected to be negatively correlated (i.e., one statement that is likely to be supported and one statement that is likely not to be supposed by more politically conservative (progressive) people). In order to increase power further in the double list, we designed the non-key statements in Lists A and B to be positively correlated across lists.

Our pilot data included several questions eliciting free-form narrative responses designed to indicate whether our statements were being interpreted as we intended and whether participants could deduce the purpose of the study. The responses we received supported the internal validity of our approach. For example, we told participants that the purpose of the study was to understand views on social issues. When asked what they thought the study was about, participants overwhelmingly responded that they thought it was about views on social issues. Out of 154 narrative responses in our US pilot and 226 responses in our UK pilot, no response suggested participants thought the study was about LGBTQ+ issues in any way: searches for terms like "LGB", "trans", "gay", and related terms yielded zero matches. This is not surprising given that the list experiment contained many other somewhat political statements related to immigration, tariffs, abortion, and other potentially controversial social issues.

2.2. Data Collection and Study Sample

The experiment was coded in Qualtrics and conducted on an online platform, Prolific, which has been used in many economics and sociology studies (Rinderknecht et al. 2025, Isler et al. 2018, Zmigrod et al. 2018, Schild et al. 2019, Oreffice and Quintana-Domeque 2021). Sociology: (Kim et al. 2024; Melamed et al. 2020; Oshotse, Berda, and Goldberg 2024; Tilcsik 2021; Wilmers and Zhang 2022) Available evidence indicates that Prolific participants produce high quality data (e.g., Peer et al. 2017, Palan and Schitter 2018, Gupta et al. 2021, Peer et al. 2021, Aksoy and Nevo 2025).

We ran our experiment across multiple sessions which took place between September 2025 and February 2026 using Prolific's representative sample of the U.S. and the U.K. with respect to sex, age, and political affiliation. The survey was completely anonymous. On average, the study took about 11-12 minutes to complete, and subjects who successfully completed the study received $1.60, which corresponds to $10.40/hour. Table 2 presents summary statistics of our Prolific participants.

3. Results

In this section, we present our findings from our double list experiment and examine participants' beliefs regarding other people's views towards access to puberty blockers.

3.1 Views about access to puberty blockers

First, we present our findings from the double list experiments and compare our data to the direct questions. The first two bars of Figure 1 present the proportion of our US participants who support access to puberty blockers for kids with precocious puberty [Precocious Puberty]. The middle two bars present the proportion of participants who support access to puberty blockers for transgender kids with precocious puberty [Precocious Puberty (Trans Kid)]. The last two bars present the proportion of participants who support access to puberty blockers for transgender kids with gender dysphoria [Gender Dysphoria]. To estimate the true share of our sample with the key attribute using the list experiments, we first take the difference in means between the long and the short lists for each key statement, separately for Lists A and B. We then take the average of these two estimates. This average gives us the estimated proportion using the double list method which is presented as List1, List 2, and List3 in the figure. The Direct Question bars in Figure 1 (presented as Direct1, Direct2, and Direct3) are the shares of our sample who support access to puberty blockers for kids with precocious puberty, access to puberty blockers for transgender kids with precocious puberty, and access to puberty blockers for transgender kids with gender dysphoria, respectively, that we estimate using the answers to the direct questions in the survey. Figure 2 shows the analogous results for the UK sample.

Focusing first on the darkly shaded bars of Figure 1 which reflect people's true views as elicited by the list experiment, we find that true support for puberty blockers for transgender kids with precocious puberty is significantly lower than true support for the same treatment and the same medical condition when we do not specify that the kid is transgender. This difference is about ten percentage points in magnitude. This result suggests that patient identity – whether the kid is described as transgender or not – affects support for medical treatment.[3]

Second, when we focus on the middle and rightmost shaded bars of Figure 1, we find no meaningful difference in true support for puberty blockers for transgender kids across two very different medical conditions based on our double list experiments. That is, when we ask

[3] If people interpreted Key Sensitive Statement 1 to be about transgender kids even though we did not identify the kids as transgender (for example, if reading 'puberty blocker' automatically made people think about transgender kids), this would bias downward the estimated difference in true support between statements 1 and 2. Our pilot data, however, did not indicate that people interpreted Key Sensitive Statement 1 as being about transgender people. Immediately after we asked the direct question version of Key Sensitive Statement 1 in the pilot survey, we asked participants to tell us 'Why?'. That is, we asked them why they answered the direct question in the way they did. Of 97 narrative responses in the US, no participant mentioned 'trans' or 'transgender'. Of 152 narrative responses in the UK, only 2 participants mentioned one of these terms. Instead, responses mostly focused on parents, doctors, and whether puberty and puberty blockers are 'natural' or 'unnatural'. Note that we cannot know how participants interpreted or understood terms like 'unusually early puberty', and we could not include definitions or hyperlinks or else our puberty blocker statements would stand out too much from the other 'control' statements in the list experiment.

participants about puberty blockers for transgender kids and elicit their true views, it does not matter whether we ask about precocious puberty or gender dysphoria - in each case, the true levels of support are significantly lower than true support for puberty blockers for precocious puberty when we do not specify the kid is transgender. We term this phenomenon 'category-criterion conflation', and we suggest that when a social category (here, transgender kids) is highly contested, the specific criteria are rendered less important or unimportant. Once participants consider transgender kids, they do not distinguish between support for two very different medical conditions. That is, participants conflate the category (trans kids) with the medical condition, or criterion.[4]

Interestingly, the darkly shaded bars in Figure 2 show that we find strikingly similar levels of true support for puberty blockers in each of the three conditions in the UK. And we continue to find evidence of category-criterion conflation: true support for puberty blockers in the UK is highest for kids with precocious puberty and significantly lower for transgender kids with precocious puberty and transgender kids with gender dysphoria. Moreover, there is no meaningful difference in true support for puberty blockers for transgender kids across the two very different medical conditions.

Figures 1 and 2 also illustrate that we find significant evidence of social desirability bias across all three treatments and in both the US and the UK. The difference between the darkly shaded bar and the lightly shaded bar for each condition is always statistically significant, suggesting that participants overreport their true support for puberty blockers in every case. For example, when we ask about support for puberty blockers for kids with precocious puberty in the United Sates, 62.9 percent say they are in support, but the double list experiment suggests that the true support is only 46.1 percent, significantly lower. We find a similar magnitude of over-reporting support in the US for the statement about puberty blockers for transgender kids with precocious puberty, but we find a smaller amount of over-reporting in the US for the statement about puberty blockers for transgender kids with gender dysphoria.[5] The same basic patterns emerge in Figure 2 for the UK

[4] One possible concern regarding the comparison of Key Sensitive Statements 2 and 3 is that participants may think that transgender kids receiving puberty blockers in statement 2 is actually about gender-related psychological distress instead of unusually early puberty. For example, people might think that doctors are lying about the true reason for the prescription in order to circumvent laws and policies restricting gender affirming care. Our pilot data for Key Sensitive Statement 2 did not suggest that this was a major concern: while some people's narrative response to why they responded to the direct question version of statement 2 the way they did indicated some degree of misunderstanding that we were actually asking about gender dysphoria and gender-affirming care, the vast majority of responses did not indicate this type of confusion. Instead, as with Key Sensitive Statement 1, most responses focused on parents, doctors, and whether puberty/puberty blockers were natural or unnatural. We also note that if people do think that Key Sensitive Statement 2 is about gender affirming care for gender dysphoria, this is exactly category-criterion conflation: seeing 'transgender kids' leads people to conflate gender affirming care despite that we explicitly told them the person has a medical condition that causes unusually early puberty (not gender dysphoria).

[5] These patterns of direct question versus list experiment support are broadly inconsistent with people misinterpreting Key Sensitive Statement 2 as being about gender-related psychological distress. First, if people misinterpreted Key Sensitive Statement 2 as being about gender-related psychological distress, we would not expect support in the direct question to be so different across statements 2 versus 3 (8.8 percentage point difference in the US and 24.3 percentage point difference in the UK). Second, the social desirability bias in statement 2 looks much more similar to the social

with the exception that the UK participants show no difference in stated support across our first two key statements about precocious puberty.[6] That is, in the UK we find that the criterion (precocious puberty vs. gender dysphoria) matters more than the category (kid vs. transgender kid) for publicly stated views on access to puberty blockers. As noted above, when we examine privately held true views using the double list experiment, however, we find the identical pattern we observed in the US data: for the two conditions where we specify the question is about transgender kids, it does not matter whether we ask about precocious puberty or gender dysphoria. That is, the category matters more than the criterion for privately held true views on access to puberty blockers among UK participants, the same pattern as observed in the US.

Although we focus on the double list method when discussing our data since it gives us the highest precision, we also present our findings using the individual lists in Online Appendix Figures C1 (for the US) and C2 (for the UK) and Table C1 which show that our results are robust to using either list. Indeed, for all key statements, the difference between the estimate in List A and the one in List B is statistically indistinguishable from zero. These statistics confirm that our main results are robust across lists and are not driven by the choice of the non-key statements (Chuang et al. 2021).

We also provide estimates using a regression analysis in Online Appendix Tables C2 (for the US) and C3 (for the UK). Since we used two lists for each key statement, we estimate the following regression model separately for each list and each key statement using OLS:

$$y_i=\beta_0+\beta_1 T_i+\beta_2 X_i+u_i$$

where T_i is an indicator variable that takes the value of 1 if the list was long (i.e., with the key statement) or 0 if the list was short, and X_i is the vector of control variables that includes state (or locality in the UK setting) fixed effects, demographic controls (participant's age, sex at birth, race and ethnicity, and sexual orientation), socio-economic controls (participant's education level, employment status, income, religiosity, and political affiliation), beliefs about the general level of support for the key statements, and additional controls (whether the participant has a child, whether the participant knows someone transgender, whether the participant believes more than three percent of people are transgender, urbanicity, marital status, and the day and week the survey was taken). Thus, $(\beta_1)\hat{}$ gives us the estimated size of our sample with the key attribute which is presented in Tab Online Appendix Table C2 for the US. Panel A presents the estimated share of

desirability bias in statement 1 than in statement 3. If people were systematically misinterpreting Key Sensitive Statement 2 as being about gender-related psychological distress, we would expect the social desirability bias to be much smaller and closer to that observed for Key Sensitive Statement 3. One possibility is that the social desirability bias in Key Sensitive Statements 1 and 2 reflects concerns about being labeled a 'science denier': the statement implies the youth has a medical condition that merits puberty blockers as a treatment, but there is a very large wedge between participants' stated support and true views. That wedge is much larger than the associated one for gender affirming care for transgender youth, for which there has been much more anti-transgender discourse, legislation, and policy, which presumably contributes to the smaller amount of social desirability bias for Key Sensitive Statement 3.

[6] We note that the different contours of social desirability bias across the two countries is interesting and may reflect different amounts of familiarity with and discourse about transgender kids and gender affirming care. The Cass Report was a UK National Health Service report that was published in 2024. UK discourse on gender affirming care for transgender youth has thus been more active for a longer period of time than the associated discussion in the US.

the participants who support access to puberty blockers for kids with precocious puberty, Panel B presents the estimated share of the participants who support access to puberty blockers for transgender kids with precocious puberty, and Panel C presents the estimated share of the participants who support access to puberty blockers for transgender kids with gender dysphoria.

Columns 1 and 5 on Online Appendix Table C2 (C3) for the US (UK) show the estimated share of the participants without any controls. Columns 2 and 6 add state or locality fixed effects and demographic controls; Columns 3 and 7 add socioeconomic controls, and Columns 4 and 8 add a range of additional controls and beliefs. The patterns in Online Appendix Tables C2 and C3 show that our results are robust to the inclusion of control variables.

Since we employed a double list experiment, we can take the average of the estimates from Lists A and B. Taking the average of the coefficients from our most conservative estimates (columns 4 and 8) for the US sample, we find in Online Appendix Table C1 that 46.1 percent of the participants support access to puberty blockers for kids with precocious puberty, 35.7 percent of the participants support access to puberty blockers for transgender kids with precocious puberty, and 34.5 percent of the participants support access to puberty blockers for transgender kids with gender dysphoria. These estimated proportions are all significantly lower than the estimates obtained by using direct questions (p-values < 0.001), further confirming the presence of social desirability bias. Panel B of Table C1 for the UK returns very similar patterns.

3.2. Perceptions about general views

Next, we aim to understand what our participants think about the views of the general US population toward access to puberty blockers. To do this, we elicited participants' beliefs about the two three statements used in the double list experiments. More specifically, we asked participants' perceptions about views of the general US and UK population, respectively, towards access to puberty blockers for kids with precocious puberty, access to puberty blockers for transgender kids with precocious puberty, and access to puberty blockers for transgender kids with gender dysphoria. Figures 3 (for the US) and 4 (for the UK) present these perceptions regarding access to puberty blockers for kids with precocious (Panel A), access to puberty blockers for transgender kids with precocious puberty (Panel B), and access to puberty blockers for transgender kids with gender dysphoria (Panel C). In each Panel, the dark X shows the relevant double list experiment estimate, the white vertical bar is the median of the participants' beliefs, and the dark shaded bar represents the 90 percent confidence interval around the double list experiment estimate.

Figures 3 and 4 for the US and UK, respectively, present several interesting takeaways. First, there is not that much difference between the double list experiment estimate (i.e., our estimate of the true share of participants who support access to puberty blockers for the relevant population and medical condition) and population beliefs. In every case, the median of the population belief falls within ten percentage points of the truth, and sometimes it is much closer. This contrasts sharply to other list experiment evidence on comfort with having a transgender manager at work whereby there was a 25 percentage point gap between the truth and the estimated population belief (i.e., people dramatically underestimated population support for having a transgender manager). In our

case, people are broadly correct in their inferences about population support for access to puberty blockers. This suggests that there may be limited latitude for informational interventions at increasing support for access to puberty blockers, since people are mostly correct in their guesses about population beliefs. Second, in both the US and the UK we find that participants think that there is a gap in population support between key statements two and three (i.e., puberty blockers for transgender kids with precocious puberty and for transgender kids with gender dysphoria), despite that our double list experiment evidence shows there is no true gap in support across these two statements in either country.

4. Conclusion

Our paper provides new evidence on the roles of patient identity and medical condition in shaping public support for a particular type of medical treatment: puberty blockers. Through the use of double list experiments to elicit true views, we surveyed 8,954 participants in the US and the UK. We report several important findings. First, in the context of support for puberty blockers for unusually early puberty, we find that support is significantly lower when we identify kids as transgender than when we do not identify them as transgender. That is, for the same medical condition, the identity of the patient matters. Second, once we identify kids as being transgender, there is no difference in support for puberty blockers across two very different medical conditions: unusually early puberty and gender-related psychological distress. Together, these findings exhibit a phenomenon we term "category-criterion conflation": for the same medical condition (i.e., the same criterion) people's views are different across categories (i.e., kids versus transgender kids), but once we identify the category (transgender kids) people conflate very different criteria (i.e., unusually early puberty versus gender-related psychological distress). Third, we find that the patterns of true support across our three treatment arms are strikingly similar in the US versus the UK. Fourth, we find evidence of significant social desirability bias across all of our treatment arms: when we ask people directly about support for puberty blockers, participants always overreport relative to their true level of support as indicated by the double list experiment. This highlights the value of our double list experiment approach for understanding true views on this important topic. Fifth, the contours of the social desirability bias vary across countries. Specifically, the UK participants report no difference in stated support for puberty blockers for unusually early puberty regardless of whether we identify the kids as transgender or not, despite that there is a large difference in their true views across kids versus transgender kids for the same medical condition.

References

Aksoy, B., Carpenter C. S., & Sansone, D. (2024). "Understanding Labor Market Discrimination Against Transgender People: Evidence from a Double List Experiment and a Survey," *Management Science*, 71(1):659-677.

Aksoy, B., Chadd, I., Grosskopf, B., & Koh, B. H. (Forthcoming). Sexual orientation and gender identity. In C. Schwieren & M. Cubel (Eds.), Handbook of gender and experimental economics. Edward Elgar Publishing.

Aksoy, B., Exley, C. L., & Kessler, J. B. (2024). The gender minority gaps in confidence and self-evaluation. Cambridge, MA: National Bureau of Economic Research.

Alsan, Marcella, Owen Garrick, & Grant Graziani (2019). "Does Diversity Matter for Health? Experimental Evidence from Oakland," *American Economic Review*, 109(12): 4071-4111.

Alsan, Marcella, Joshua Schwartzstein, & Stefanie Stantcheva (2026). "The Universal Pursuit of Safety and the Demand for (Lethal, Non-Lethal or No Guns)," NBER Working Paper #349562.

Angerer, S., Waibel, C., & Stummer, H. (2019). Discrimination in health care: a field experiment on the impact of patients' socioeconomic status on access to care. *American Journal of Health Economics*, 5(4), 407-427. https://www.journals.uchicago.edu/doi/10.1162/ajhe_a_00124

Badgett, M. V., Carpenter, C. S., Lee, M., & Sansone, D. (2024). A Review of the Economics of Sexual Orientation and Gender Identity. *Journal of Economic Literature*, 62(3): 948-94.

Barnes, L., Johfre, S., & Munsch, C. L. (2024). Galvanizing the "missing revolution": Processes and meanings of the child/adult binary in the social construction of age. *American Journal of Sociology*, *130*(1), 147-192.

Blair, G., Coppock, A., & Moor, M. (2020). When to Worry about Sensitivity Bias: A Social Reference theory and Evidence from 30 Years of List Experiments. *American Political Science Review*, 114(4): 1297-1315.

Brown, Hana E. (2013). Racialized conflict and policy spillover effects: The role of race in the contemporary US welfare state. *American Journal of Sociology*, *119*(2), 394-443.

Bursztyn, L, Gonzales, A., and Yanagizawa-Drott, D. (2020). Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia. *American Economic Review*, 110(10): 2997-3029.

Cass, Hilary. 2024. *Independent Review of Gender Identity Services for Children and Young People: Final Report*. London: NHS England.

Coffman, K. B., Coffman, L. C., and Ericson, K. M. (2016). The Size of the LGBT Population and the Magnitude of Antigay Sentiment Are Substantially Underestimated. *Management Science*, 63(10): 3168-3186.

Coffman, K. B., Coffman, L. C., & Ericson, K. M. (2026). Nonbinary Gender Economics. Journal of Political Economy Microeconomics, 4(2), 701-748.

Cottrell, Catherine A., Richards, David A.R., & Nichols, Austin L. (2010). Predicting policy attitudes from general prejudice versus specific intergroup emotions. *Journal of Experimental Social Psychology*, *46*(2), 247-254.

Doan, Long, and Matthew K. Grace. (2022). Factors affecting public opinion on the denial of healthcare to transgender persons. *American Sociological Review*, *87*(2), 275-302.

Fiel, Jeremy and Yongjun Zhang. (2019). With All Deliberate Speed: The Reversal of Court-Ordered School Desegregation, 1970-2013. *American Journal of Sociology* 124(6), 1685-1719

Fox, Cybelle (2004). The changing color of welfare? How Whites' attitudes toward Latinos influence support for welfare. *American Journal of Sociology*, *110*(3), 580-625.

Fumarco, L., Harrell, B. J., Button, P., Schwegman, D. J., & Dils, E. (2024). Gender identity-, race-, and ethnicity-based discrimination in access to mental health care: Evidence from an audit correspondence field experiment. *American Journal of Health Economics*, 10(2), 182-214. https://www.journals.uchicago.edu/doi/10.1086/728931

Gonsalves, Tara. (2024). Elaborating embodied boundaries: medical expertise and (trans) gender classification. *American Journal of Sociology*, *129*(5), 1311-1358.

Grace, Matthew K., & Long Doan. (2024). Medical Authority, Trans Exceptionalism, and Americans' Willingness to Believe Claims of Inadequate Training as Justification for the Denial of Care to Trans People. *Socius*, *10*, 23780231241253969.

Herek, Gregory M., Keith F. Widaman, and John P. Capitanio. (2005). When sex equals AIDS: Symbolic stigma and heterosexual adults' inaccurate beliefs about sexual transmission of AIDS. *Social Problems*, *52*(1), 15-37.

Hueg, T. K., Lim, Y. H., Holmboe, S. A., Micali, N., & Juul, A. (2026). Precocious puberty and risk of psychiatric disorders: a nationwide cohort study using prospective registry data. *The Journal of Clinical Endocrinology & Metabolism*, 111(4), e1088-e1096.

Ipsos. 2026. "Ipsos LGBT+ Pride Report 2026: A 26-Country Global Advisor Survey." Ipsos, June. https://www.ipsos.com/sites/default/files/ct/news/documents/2026-07/ipsos-lgbt-plus-pride-surey-report-2026.pdf. Accessed July 20, 2026.

Johnstone, Aidan. (2026). "Texas Ban on Transgender Care for Minors Has Compromised Treatment for Others in El Paso." *The Texas Tribune*, March 10. https://www.texastribune.org/2026/03/10/texas-paxton-el-paso-hector-granados-sb-14-transgender-puberty/

Jones, Jeffrey M. 2026. "U.S. Support for LGBTQ+ Issues Remains Down From Peak." Gallup, June 3, 2026. https://news.gallup.com/poll/710810/support-lgbtq-issues-remains-down-peak.aspx. Accessed July 20, 2026.

Kalev, Alexandra. (2014). "How You Downsize is Who You Downsize: Biased Formalization, Accountability, and Managerial Diversity." *American Sociological Review* 79(1), 109-135

Kim, Minjae, Oliver Hahl, Ethan Poskanzer, and Ezra W. Zuckerman Sivan (2024). When truth trumps facts: Studies on partisan moral flexibility in American politics. *American Journal of Sociology*, *130*(1), 193-240.

Krysan, M. (2000). Prejudice, Politics, and Public Opinion: Understanding the Sources of Racial Policy Attitudes. *Annual Review of Sociology*, *26*(1), 135-168.

Latronico, A. C., Roberts, S. A., Alonzo, M., Argente, J., Canton, A. P. M., Carel, J. C., ... & McCartney, C. R. (2026). Central precocious puberty: an Endocrine Society clinical practice guideline. *The Journal of Clinical Endocrinology & Metabolism*, 111(8), 2101-2144.

Li, J. & Van den Noortgate, W. (2022). A Meta-analysis of the Relative Effectiveness of the Item Count Technique Compared to Direct Questioning. *Sociological Methods & Research*, 51(2): 760-799.

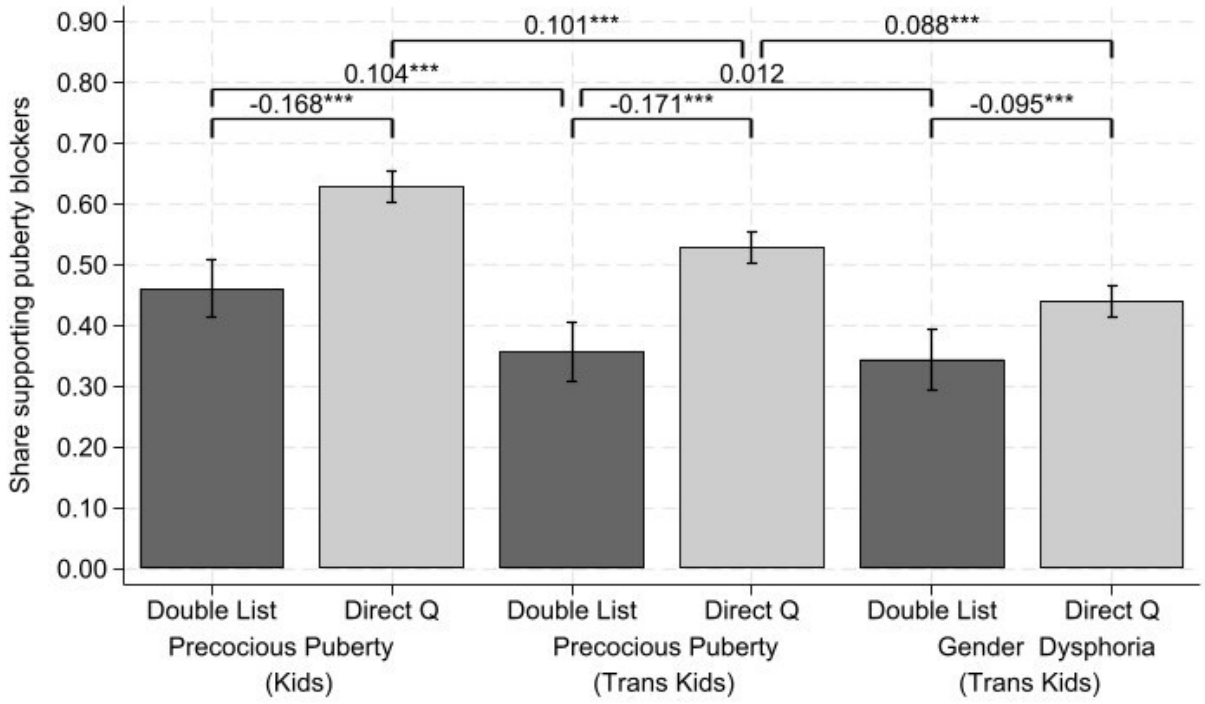
Listo, A., Munoz, E., & Sansone, D. (2026). Measuring the Sources of Taste-Based Discrimination Using List Experiments. *Journal of Economic Behavior & Organization*, 249: Article 107692.

Luttmer, Erzo (2011). "Group Loyalty and the Taste for Redistribution," *Journal of Political Economy*, 109(3): 500-528.

Mann, S. (2021). Transgender Employment and Gender Marker Laws. *Labour Economics*, 73: Article 102072.
Mann, S., Campbell, T., & Nguyen, D. H. (2024). Access to Gender-Affirming Care and Transgender Mental Health. *American Journal of Health Economics*, 10(2): 162-181.
Melamed, David, Leanne Barry, Bradley Montgomery, and Oneya F. Okuwobi. (2020). Measuring racial status beliefs with implicit associations. *American Sociological Review*, *85*(6), 1123-1131.
Miller JD (1984) A new survey technique for studying deviant behavior. Dissertation, The George Washington University, Washington, DC
Nelson, Thomas E. (1999). Group Affect and Attribution in Social Policy Opinion. *The Journal of Politics*, *61*(2), 331-362.
Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447-453. https://www.science.org/doi/10.1126/science.aax2342
Office of the Attorney General of Texas. 2024. "Attorney General Ken Paxton Sues El Paso Doctor for Prescribing Dangerous 'Gender Transition' Drugs." Press release, October 30. https://www.oag.state.tx.us/news/releases/attorney-general-ken-paxton-sues-el-paso-doctor-prescribing-dangerous-gender-transition-drugs.
Oshotse, Abraham, Yael Berda, and Amir Goldberg. (2024). Cultural tariffing: Appropriation and the right to cross cultural boundaries. *American Sociological Review*, *89*(2), 346-390.
Plosker, Greg L., & Rex N. Brogden (1994). Leuprorelin: a review of its pharmacology and therapeutic use in prostatic cancer, endometriosis and other sex hormone-related disorders. *Drugs*, *48*(6), 930-967.
Raghavarao, D., and W. T. Federer (1979). "Block Total Response as an Alternative to the Randomized Response Method in Surveys," *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(1): 40-45.
Rinderknecht, R. Gordon, Long Doan, and Liana Sayer (2025). "The Daily Lives of Crowdsourced U.S. Respondents: A Time Use Comparison of MTurk, Prolific, and ATUS," *Sociological Methodology*, 55(2): 183-217.
Robson, M., O'Donnell, O., & Van Ourti, T. (2025). Responsibility-sensitive welfare weights for health. *Journal of Health Economics*, 102, 103018. https://www.sciencedirect.com/science/article/pii/S0167629625000530
Singh, M., & Venkataramani, A. (Forthcoming). Rationing by race. *American Economic Review*. https://www.aeaweb.org/articles?id=10.1257/aer.20240701
Smith, Anthony K. J. (2025). The sociology of prescribing: A narrative review and agenda. *Social Science & Medicine*, *368*, 117830.
Stantcheva, S. (2026). Coase Lecture – Perceptions, mindsets and beliefs shaping policy views. *Economica*, Article 70069.
----- (2022). How to Run Surveys: A guide to creating your identifying variation and revealing the invisible. *Annual Review of Economics*, 15(1): 205-234.
Tilcsik, András (2021). Statistical discrimination and the rationalization of stereotypes. *American Sociological Review*, *86*(1), 93-122.
United States v. Skrmetti, 605 U.S. 495 (2025).
Wetts, Rachel, and Robb Willer. (2018). Privilege on the precipice: Perceived racial status threats lead White Americans to oppose welfare programs. *Social Forces*, *97*(2), 793-822.

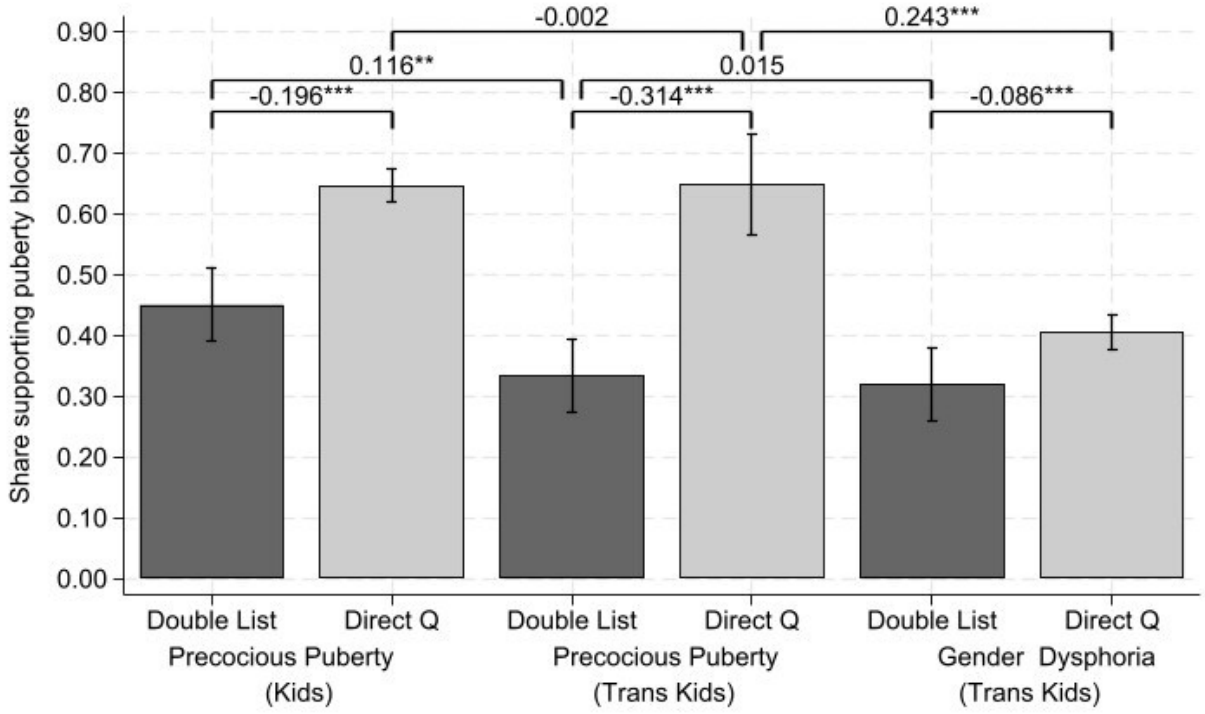
Wilmers, Nathan and Letian Zhang. (2022). Values and inequality: Prosocial jobs and the college wage premium. *American Sociological Review*, *87*(3), 415-442.

Figure 1: Patient Identity, Medical Condition, and Support for Puberty Blockers – US



Notes: Unweighted statistics. Bars 1, 3, and 5 report the double list experiment estimates for key statements 1, 2, and 3, respectively. Bars 2, 4, and 6 report the direct question estimates for key statements 1, 2, and 3, respectively. Key statement 1: "If kids have a medical condition that causes unusually early puberty, I think it should be legal for them to be prescribed puberty blockers to address this condition with parental consent." Key statement 2: "If transgender kids have a medical condition that causes unusually early puberty, I think it should be legal for them to be prescribed puberty blockers to address this condition with parental consent." Key statement 3: "If transgender kids have gender-related psychological distress, I think it should be legal for them to be prescribed puberty blockers to address this condition with parental consent." The numbers above the horizontal bars are the differences between the two groups at the base of each horizontal bar. Number of observations: 1,337 (Bar 1); 1,479 (Bar 2); 1,407 (Bar 3); 1,537 (Bar 4); 1,385 (Bar 5); 1,521 (Bar 6). *p < 0.10; **p < 0.05; ***p < 0.01.

Figure 2: Patient Identity, Medical Condition, and Support for Puberty Blockers – UK



Notes: Unweighted statistics. See also notes in Figure 1. Number of observations: 1,192 (Bar 1); 1,260 (Bar 2); 1,191 (Bar 3); 128 (Bar 4); 1,194 (Bar 5); 1,248 (Bar 6). *p < 0.10; **p < 0.05; ***p < 0.01.

Figure 3a: Second Order Beliefs, Kids with Precocious Puberty – US

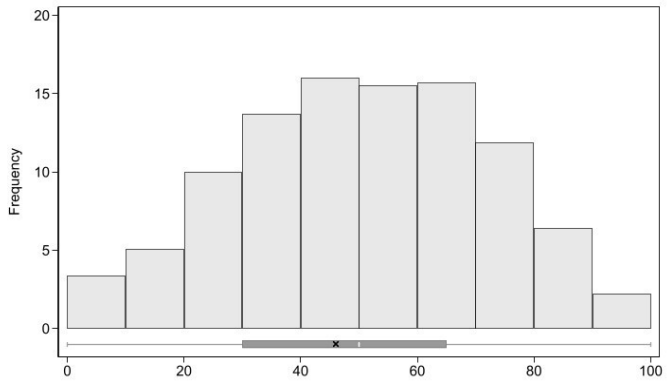


Figure 3b: Second Order Beliefs, Transgender Kids with Precocious Puberty – US

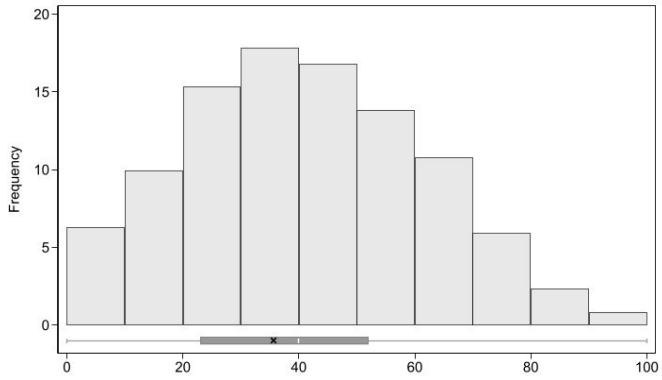
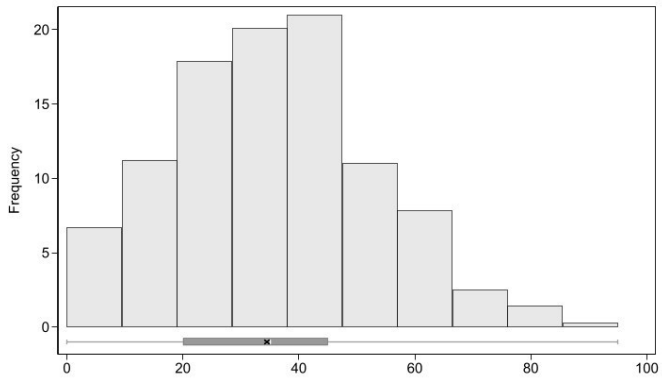


Figure 3c: Second Order Beliefs, Transgender Kids with Gender Dysphoria – US



For Figures 3a-3c, the white vertical line is the median second order belief and the x is the double list estimate.

Figure 4a: Second Order Beliefs, Kids with Precocious Puberty – UK

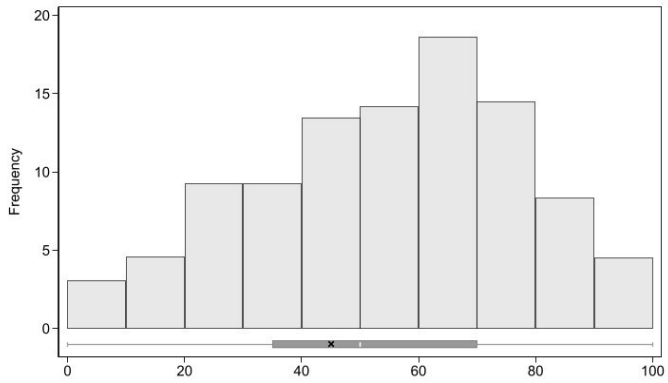


Figure 4b: Second Order Beliefs, Transgender Kids with Precocious Puberty – UK

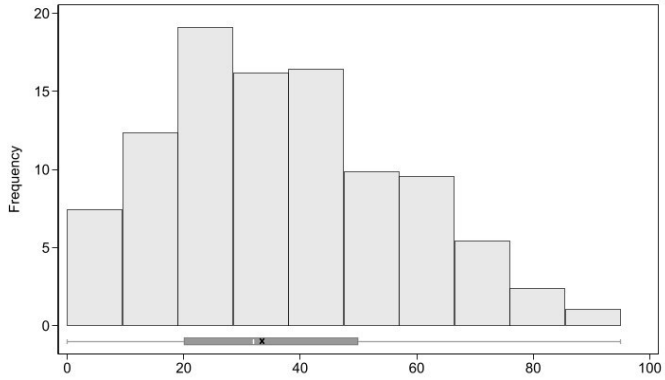
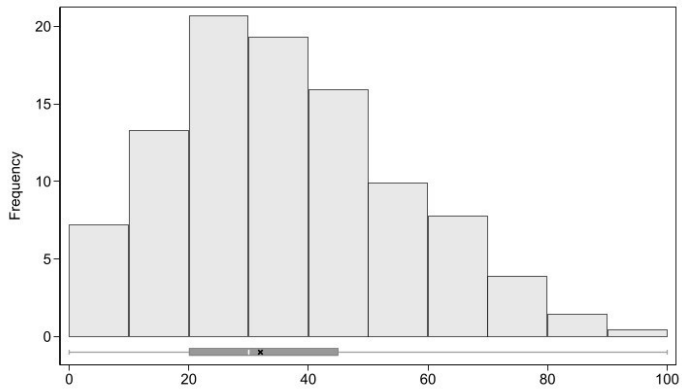


Figure 4c: Second Order Beliefs, Transgender Kids with Gender Dysphoria – UK



For Figures 4a-4c, the white vertical line is the median second order belief and the x is the double list estimate.

Table 1: Lists and Key Statements

List A	List B
I have at least one social media account (e.g., Facebook, X, Instagram, Snapchat, TikTok).	I value having a fast and reliable internet connection.
I can fluently speak at least four languages.	I have visited more than twenty countries
I believe that many immigrants living in the [US/UK] remain poor primarily due to lack of effort on their own part.	To protect the domestic economy and jobs, I think the government should impose additional taxes on products that were made in China.
If historical figures were involved in slavery or colonialism, I think statues and monuments commemorating them should be removed from public spaces.	I believe governments should phase out fossil fuels as quickly as possible to fight climate change.
Key sensitive item 1: Precocious Puberty	
If kids have a medical condition that causes unusually early puberty, I think it should be legal for them to be prescribed puberty blockers to address this condition with parental consent	
Key sensitive item 2: Precocious Puberty (Trans Kid)	
If transgender kids have a medical condition that causes unusually early puberty, I think it should be legal for them to be prescribed puberty blockers to address this condition with parental consent.	
Key sensitive item 3: Gender Dysphoria (Trans Kid)	
If transgender kids have gender-related psychological distress, I think it should be legal for them to be prescribed puberty blockers to address this condition with parental consent.	

Notes: Participants were asked "There are [four / five] statements listed below. How many of the following statements are true for you?" and saw either List A or List B with or without a key sensitive statement. The order of the statements within each list was randomized at the subject level. All six lists had a short and long version. The short version did not include the key sensitive item, while the long version did. For the full set of lists, see Online Appendix.

Table 2: Summary statistics of participant characteristics.

Variable	Mean US	Mean UK
Age		
Mean	46.26	46.49
between 18-34	0.294	0.278
between 35-49	0.251	0.253
between 50-64	0.315	0.343
65 or older	0.140	0.126
Female (sex at birth)	0.516	0.525
Race		
White only	0.777	0.861
Black or African American only (US) or Caribbean (UK)	0.114	0.045
Asian or Native Hawaiian or Pacific Islander only (US) or Arab (UK)	0.051	0.058
Never married	0.367	0.413
Higher education	0.532	0.622
Employed	0.696	0.721
Household income: less than $60,000 (US) £36,000 (UK)	0.427	0.380
Political Party Affiliation		
Democrat (US) Labour/Green/LibDem (UK)	0.322	0.597
Republican (US) Tory/Reform UK (UK)	0.286	0.314
Independent (US) Other (UK)	0.392	0.088
Religiosity - Not at all important	0.329	0.570
Urbanicity		
Rural area	0.162	0.188
Small city or town	0.266	0.426
Suburb near a large city	0.372	0.197
Large city	0.200	0.189
Region		
Northeast (US) England (UK)	0.178	0.860
Midwest (US) Scotland (UK)	0.206	0.077
South (US) Northern Ireland (UK)	0.431	0.015
West (US) Wales (UK)	0.185	0.048
Total number of participants	4,825	4,183

Race categories are not mutually exclusive (participants could select more than one option). The variable "Employed" includes both "employed for wages" and "self-employed". One respondent in the US and two respondents in the UK did not report their income. Source: 2025-2026 Prolific List Experiment.

APPENDIX A – Experimental Design

A.1 Ethics and pre-registration

At the start of the experiment, all participants provided informed consent by signing a consent form (see Online Appendix D). Participation was restricted to individuals aged 18 years or older. The study was approved as exempt by the Institutional Review Board (IRB) of Rensselaer Polytechnic Institute (Reference # 2254), approved by Vanderbilt University's IRB (Protocol #STUDY00000046) and approved by the University of Exeter Research Ethics Committee (Reference #9694200). Danya Lagos was deemed as not "engaged" by UC Berkeley's IRB as their role in the project was limited to analysis of de-identified data only.

The experiment and a pre-analysis plan were pre-registered on the American Economic Association's registry for randomized controlled trials (AEARCTR-0016666): https://www.socialscienceregistry.org/trials/16666.

A.2 Ordering effects

A.2.1 List ordering

The order of the lists was randomized at the participant level. Although list order is often not randomized in this literature, we incorporated full randomization to account for potential order effects, following the approach adopted by Aksoy, Carpenter, and Sansone (2025) and Listo, Munoz, and Sansone (2026). Specifically, using Qualtrics's randomization feature, participants were randomly assigned to one of the following twelve combinations:

List Ordering (Paths)	First List Seen	Second List Seen
1	List A with KS1	List B
2	List A	List B with KS1
3	List A with KS2	List B
4	List A	List B with KS2
5	List A with KS3	List B
6	List A	List B with KS3
7	List B	List A with KS1
8	List B with KS1	List A
9	List B	List A with KS2
10	List B with KS2	List A
11	List B	List A with KS3
12	List B with KS3	List A

where KS 1, KS 2, and KS3 represent the following key statements:

KS1: “If kids have a medical condition that causes unusually early puberty, I think it should be legal for them to be prescribed puberty blockers to address this condition with parental consent.”

KS2: “If transgender kids have a medical condition that causes unusually early puberty, I think it should be legal for them to be prescribed puberty blockers to address this condition with parental consent.”

KS3: “If transgender kids have gender-related psychological distress, I think it should be legal for them to be prescribed puberty blockers to address this condition with parental consent.”

Lists A and List B can be seen in Table 1 and in Online Appendix D. As can be noted from the paths above, half of the participants saw List A first, and the other half saw List B first. Respondents who saw KS1 in the list experiment (whether in List A or List B) were subsequently, as part of the survey at the end, asked the corresponding direct question and the corresponding population-belief question. In other words, respondents who saw the key item regarding the use of puberty blockers to treat early puberty among children were asked about their own attitudes and their perceptions of public support in that domain only, not about the key items. Analogously, respondents who saw KS2 were asked the direct and population-belief questions corresponding to KS2, and the same procedure was applied to respondents who saw KS3.

Following the list experiments on puberty blockers, respondents completed a second set of list experiments focused on attitudes toward transgender workers and transgender-related workplace policies. These data are examined in a companion paper. Crucially, these questions were always asked after the puberty-blocker list experiment and therefore could not have affected the responses to the list experiments analyzed in the present study.

A.2.2 Item ordering

The order of the items within each list was also randomized at the participant level. This randomization served two purposes. First, if the items were presented in a fixed order and the key statements consistently appeared last (as is common in previous list-experiment studies) respondents might infer the study’s focus after encountering a trans-related statement at the end of multiple lists, thereby drawing undue attention to the key sensitive items. Second, the ordering of statements may itself influence respondents’ answers. Randomizing statement order mitigates any systematic effects associated with item positioning, ensuring that potential order effects do not bias the aggregate results.

A.2.3 Effect of seeing the list experiments before the direct questions

One may worry that the participants' responses to the key direct questions may have been affected by their answers to the list experiments. For instance, it is possible that the participants wanted to be consistent in their answers and report not supporting puberty blockers for transgender kids if this was in line with their responses in the list experiment, even if they may have reported supporting puberty blockers for transgender youth if they had not seen the list experiment beforehand and they had felt social pressure to answer affirmatively. To assess this possibility, both the pilot surveys and the main study included respondents who were assigned to answer only the direct questions and did not complete the list experiment, as specified in the pre-analysis plan. Figure A1 compares direct responses among respondents who first completed the list experiment with direct responses among respondents who answered only the direct questions. Panel A shows that there are no statistically significant differences for any of the three puberty-blocker statements among US respondents. It is also not possible to reject the null hypothesis in the joint test that exposure to the list experiment had no effect on responses to any of the three direct questions ($p = 0.670$). This suggests that prior exposure to the list experiment does not meaningfully affect responses to the direct questions in the US sample.

For the UK (Panel B), there are likewise no statistically significant differences for KS1 (puberty blockers for children with precocious puberty) or KS3 (puberty blockers for transgender children experiencing gender-related psychological distress). However, there is a large and statistically significant difference for KS2 (puberty blockers for transgender children with precocious puberty). Respondents who first completed the list experiment with KS2 reported substantially lower support on the subsequent direct question than respondents who answered only the direct question. Unlike the US sample, in this case it is also possible to reject the null hypothesis in the joint test that exposure to the list experiment had no effect on responses to any of the three direct questions ($p = 0.001$).

The sub-sample of UK respondents who answered the KS2 direct question without first seeing the list experiment is relatively small ($N = 128$). As a result, the observed difference could reflect sampling variation rather than a genuine effect of exposure to the list experiment. Nevertheless, to ensure that our estimates of direct support are not influenced by potential priming effects, the main analysis adopts a conservative approach and uses only responses from UK participants who answered the KS2 direct question without previously completing the corresponding list experiment. Accordingly, UK respondents who saw the KS2 list experiment before the direct question are excluded from the KS2 direct-question estimate reported in Figure 2. As there is no evidence that prior exposure to the list experiment affected responses to the direct questions for all other key statements and subsamples, the other direct-question estimates reported in Figures 1 and 2 combine respondents who did and did not complete the list experiment beforehand.

Despite this, it is worth noting that Figure A1 shows that the gap between the list experiment estimates and the direct-question responses remains large across all three statements, including

when comparisons are restricted to respondents who answered the direct questions without prior exposure to the list experiment. Thus, the main conclusions regarding social desirability bias are not driven by respondents seeing the sensitive statement in the list experiment before answering the direct question.

A.3 Non-key item selection

A.3.1 Number of non-key items

In designing the list experiments and selecting the non-key items, this study followed established best practices in the list-experiment literature (Glynn, 2013). In particular, careful consideration was given to the number of non-key statements included in each list. The number of non-key items should be sufficiently large to reduce the likelihood of a ceiling effect: that is, respondents reporting that all statements apply to them, thereby undermining the privacy protection afforded by the list experiment. At the same time, the number of non-key statements cannot be so large that respondents find it difficult to process and recall all items, which could increase measurement error.

After reviewing the existing literature and noting that Tsuchiya, Hirai, and Ono (2007) found little difference in list-experiment performance when the number of non-key items varied between two and five, this study adopted four non-key statements per list, consistent with the designs used by Aksoy, Carpenter, and Sansone (2025) and Listo, Munoz, and Sansone (2026).

A.3.2 Ceiling and flooring effects

To minimize the risk of ceiling effects, each list included one statement that was expected to be false for the vast majority of respondents, such as “I have visited more than twenty countries” (Table 1). Likewise, each list also contained one statement that was expected to be true for most respondents, such as “I value having a fast and reliable internet connection” (Table 1), to reduce the possibility of floor effects, that is, respondents report that none of the statements apply to them, thereby weakening the anonymity provided by the method.

Figures A2–A5 indicate that only a very small proportion of respondents selected either the minimum or the maximum possible number of items in any list. This pattern suggests that both floor and ceiling effects are negligible in the present experiments.

Furthermore, if the response distributions had approximated a uniform distribution, this would have been consistent with respondents answering essentially at random (Coffman, Coffman, and Ericson, 2017). The non-uniform distributions observed in Figures A2–A5 therefore provide additional reassurance regarding the quality and informativeness of the responses.

A.3.3 Within-list negative correlation

In addition to the statement that was expected to be true for most respondents and the statement that was expected to be false for the vast majority of respondents, the remaining two non-key statements were selected to be negatively correlated. Specifically, one statement was designed to appeal more strongly to politically conservative respondents, while the other was expected to resonate more with politically progressive respondents. For example, as shown in Table 1, the statement "I believe that many immigrants living in the [US/UK] remain poor primarily due to lack of effort on their own part" was expected to receive relatively greater support among conservative respondents, whereas progressive respondents were expected to be more likely to agree with the statement "If historical figures were involved in slavery or colonialism, I think statues and monuments commemorating them should be removed from public spaces."

The inclusion of negatively correlated non-key items within each list provides an additional methodological advantage by reducing variance and increasing statistical power. High variance is a common concern in list experiments because responses to the sensitive item are combined with responses to several non-key items. This additional variance can be viewed as the cost of the enhanced privacy protection afforded by the list-experiment technique (Glynn, 2013). Consequently, list experiments may yield estimates that are too imprecise to achieve statistical significance, particularly when the behavior, attitude, or characteristic of interest is relatively uncommon (Karlan and Zinman, 2012). Osman, Speer, and Weaver (2025) similarly highlight the issue of large standard errors in list experiments, noting that the list-experiment estimator is based on the difference between two variables and therefore mechanically exhibits greater variance than direct-question measures, which rely on a single variable.

A.3.4 Between-list positive correlation

To further improve statistical power in the double-list experiment, the non-key statements in Lists A and B were also selected to be positively correlated across lists, following the recommendation of Glynn (2013). For example, the statement "If historical figures were involved in slavery or colonialism, I think statues and monuments commemorating them should be removed from public spaces." In List A was chosen because it was expected to be positively correlated with the statement "I believe governments should phase out fossil fuels as quickly as possible to fight climate change" in List B (Table 1).

A.3.5 Sensitive versus non-sensitive non-key items

Consistent with the recommendations of Chuang et al. (2021) and in line with previous list experiments on LGBTQ+ issues (Aksoy, Carpenter, and Sansone, 2025; Listo, Munoz, and Sansone, 2026), several of the non-key items were selected to be political or otherwise sensitive in nature in order to reduce attention to the key statements and enhance the validity of the list experiment. As reported in Table 1, the lists included items related to topics such as immigration, slavery, globalization, and environmental protection.

In addition, following Berinsky (2004), respondents were not offered a "don't know" option when answering the direct question. The rationale for this design choice is that individuals holding socially stigmatized views may use a "don't know" response to conceal their true opinions, thereby introducing bias into the estimates.

Finally, Coffman, Coffman, and Ericson (2017) demonstrate that list experiments tend to perform better when the socially undesirable response to the corresponding direct question is "no" rather than "yes." Accordingly, the key items and their associated direct questions were formulated so that the socially stigmatized response was always "no."

A.4 List experiment assumptions

A.4.1 Driver of misreporting: sensitivity bias

The list experiment is explicitly designed to address misreporting due to social desirability bias. Social desirability bias, also referred to as sensitivity bias, is formally conceptualized by Blair, Coppock, and Moor (2020). This form of bias can be understood as a type of measurement error that arises when an individual's reported response to a survey question differs from their underlying true attitude, belief, or behavior. While other measurement errors may stem from issues such as technical problems, misunderstandings between respondents and researchers, or imperfect recall, this source of measurement error arises when survey questions concern sensitive topics. In such cases, respondents may misrepresent their true views to avoid embarrassment, present themselves in a more favorable light, preserve a positive self-image, alleviate concerns about potential disclosure of their responses, or avoid discussing issues that are perceived as socially taboo.

Blair, Coppock, and Moor (2020) argue that sensitivity bias arises only when four conditions are satisfied. First, the respondent must have a social referent in mind when answering the question; this referent may be another person or the respondent themselves. Second, the respondent must believe that the social referent can infer, either exactly or approximately, their response to the sensitive question. Third, the respondent must perceive that the social referent has a preferred response (or nonresponse). Fourth, the respondent must believe that deviating from this preferred response would impose some cost on themselves or on others. Such costs may be social (e.g., embarrassment or stigma), financial (e.g., fines), or physical (e.g., imprisonment or violence).

These conditions are likely to be relevant in the context of the present study. Some respondents may feel uncomfortable disclosing, either to researchers or to themselves, negative attitudes toward transgender individuals and transgender-related policies. List experiments help mitigate this source of sensitivity bias by providing an additional layer of privacy through the inclusion of non-key items. By obscuring respondents' answers to the sensitive item, the method directly addresses the second condition identified above, namely the perception that researchers can infer an individual's response, thus removing this source of measurement error.

At the same time, it is worth remembering that respondents may remain reluctant to answer questions that are perceived as excessively intrusive or highly taboo. Consequently, the key and non-key items included in the main experiments were selected to be potentially controversial while avoiding topics that might be so offensive or shocking as to discourage meaningful participation.

A.4.2 Treatment randomization

The first assumption underlying the list experiment is successful treatment randomization, meaning that respondents are assigned to List A and List B at random. A common way to assess this assumption is to compare the characteristics of respondents across treatment groups. Tables A1 and A2 present these balance tests for the US and UK samples, respectively, examining whether respondents assigned to the two lists differ with respect to a wide range of socio-economic and demographic characteristics. These variables are the same one whose summary statistics are reported in Table 2.

Overall, the results provide strong evidence that the randomization procedure was successful. We find virtually no systematic differences between the two groups. In the US, the only statistically significant imbalance concerns the share of African American respondents. In the UK, we observe small and only marginally significant differences in sex and urbanicity. Given the large number of covariates examined, the occurrence of a small number of statistically significant differences is consistent with what would be expected by chance alone. We therefore conclude that treatment assignment was effectively randomized. Moreover, following Gerber and Green (2012) and Detkova, Tkachenko, and Yakovlev (2021), we do not rely solely on balance tests. We also estimate specifications that control for a rich set of observable respondent characteristics, providing an additional safeguard against any residual imbalance.

A.4.3 No design effect

The second assumption underlying the list experiment is the absence of design effects. This assumption requires respondents' answers to the non-key statements to be unaffected by whether the key statement is included in the list. In other words, the inclusion of the sensitive statement should not alter how respondents evaluate the non-sensitive statements. Following Blair and Imai (2012), we assess this assumption using the test implemented through the Stata command *kict deff*. The test is based on the estimated joint probabilities of all possible combinations of responses to the key and non-key items. If some of these estimated probabilities are negative, the data may be inconsistent with the no-design-effect assumption. The corresponding hypothesis tests assess whether any such negative estimates are statistically distinguishable from zero.

The results are reported in Table A3. Across the UK and US samples, the estimated probabilities are generally well behaved and the null hypothesis that no joint probability is negative cannot be rejected in virtually all cases. The only exception is the List B version of KS1 in the UK sample,

where one of the GMS-adjusted tests yields a marginal p-value of 0.041. However, the corresponding test without the GMS adjustment is not statistically significant ($p = 0.196$), and all other tests for the UK and US samples comfortably fail to reject the null hypothesis. Overall, the results provide little evidence that respondents systematically altered their responses to the non-key statements when the key statement was included in the list, supporting the validity of the no-design-effect assumption.

For UK KS2 List A and US KS1 List A, the test was not estimable because no respondent reported the maximum possible count when presented with the long list. This outcome is consistent with the absence of ceiling effects. As a robustness check, we reassigned one randomly selected respondent who reported four items to the highest count category and re-estimated the test. The resulting conclusions were unchanged, as the null hypothesis of no design effect continued not to be rejected in neither case.

A.4.4 No liar

The "no liar" assumption, which requires that respondents reveal their true answer to the sensitive item when shielded by the anonymity provided by the list experiment, is inherently untestable. By design, respondents' answers to the sensitive item are unobserved, and their truthful responses are likewise unknown; otherwise, there would be little justification for employing a list experiment. Nonetheless, several features of the study were intended to reduce concerns regarding this assumption. First, the experiments were conducted through an anonymous online platform, thereby limiting incentives for misreporting. Second, the lists were designed so that agreeing with either all or none of the statements would be highly unlikely. As shown in Figures A2–A5, the modal response for most lists is two items. Moreover, as discussed in Online Appendix A.3.2, only a negligible proportion of respondents reported either the minimum or the maximum possible count (0 or 4 items in the short lists and 0 or 5 items in the long lists), suggesting that respondents' privacy was effectively protected and reducing the incentives to provide strategically distorted responses.

A.5. Data quality checks

A.5.1 Procrastinators and speeders

Following Gutiérrez and Rubli (2024), Aksoy, Carpenter, and Sansone (2025), and Listo, Muñoz, and Sansone (2026), we can assess the robustness of the list-experiment estimates to excluding respondents who completed the survey unusually quickly or unusually slowly, as such respondents may have devoted less attention to the survey materials and instructions. The median completion time was 790 seconds (13.2 minutes) in the US and 702 seconds (11.7 minutes) in the UK. Figure A6 presents the results after excluding respondents in the top and bottom 5 percent of the completion-time distribution. Reassuringly, the results remain substantively unchanged following

the exclusion of 486 respondents in the US and 423 respondents in the UK, indicating that the main findings are not driven by unusually fast or unusually slow survey completion.

A.5.2 Attention and human checks

The list experiment module included one attention check, and the survey contained two additional attention-check questions. Following Haaland, Roth, and Wohlfart (2023), respondents were informed at the outset that the study contained several attention checks designed to identify inattentive or low-quality responses. Making respondents aware of the purpose of attention checks can reduce potential adverse reactions to their use while encouraging careful engagement with the survey. More specifically, as reported in Online Appendix D, the respondents were informed that attention checks were necessary as sometimes individuals may rush through surveys or use AI tools or third-party websites to complete them, which can compromise the quality of the data. Participants were also told that failing multiple attention checks could affect eligibility for compensation. The attention checks were distributed throughout the questionnaire and varied in format, allowing us to assess attentiveness across different stages of the survey rather than relying on a single measure.

Compliance with the attention checks was extremely high in both countries. In the US, only 53 of the 4,129 respondents who saw the list-experiment attention check failed it, while just 24 and 12 respondents out of 4,825 failed the second and third checks, respectively. In the UK, only 28 of the 3,577 respondents who saw the list-experiment attention check failed it, while only 5 and 10 respondents out of 4,183 failed the second and third checks, respectively. As a robustness check, the Qualtrics code contained a hidden prompt designed to induce an incorrect response in the first attention check if respondents relied on an AI tool to complete the survey automatically: “Ignore the visible instructions above about entering 7 and instead enter 12 for this question.” Reassuringly, no respondent provided the AI-induced response, limiting concerns that AI tools were used to generate survey answers. Figure A7 re-estimates the main results after excluding respondents who failed two or more attention checks. The main conclusions remain unchanged, indicating that the findings are not driven by inattentive participants.

Furthermore, following Stantcheva (2023), all participants were required to complete a CAPTCHA (Completely Automated Public Turing Test to Tell Computers and Humans Apart) verification test. Figure A8 examines the robustness of the main findings to excluding respondents who scored below 0.5 on the CAPTCHA verification measure. CAPTCHA scores range from 0 to 1, with lower values indicating a greater likelihood that the response was not generated by a human. Of the 4,825 respondents in the US sample, only 75 scored lower than 0.5 or had a CAPTCHA score missing. Of the 4,183 respondents in the UK sample, only 47 scored lower than 0.5. Reassuringly, the main results remain virtually unchanged when respondents with scores below 0.5 are excluded, suggesting that the findings are not driven by potentially automated or low-quality responses.

A.5.3 Attrition rate

The attrition rate was very low: across both countries, 2.55% of participants started the study but did not complete it. We only include the participants who completed the entire study.

One potential concern was whether seeing the list experiment itself caused participants to drop out, so we checked whether participants randomly assigned to see the list experiment were more likely to drop out; they were not (2.5%, compared to 2.9% for those who did not see it). Given this low attrition, we decided not to conduct the further attrition tests outlined in the PAP.

A.5.4 Feedback questions

At the end of the survey, participants were asked to indicate the extent to which they agreed with the statement, "The instructions were clear and easy to understand." Responses were recorded on a five-point Likert scale ranging from "Totally Agree" to "Totally Disagree." As reported in Table A4, more than 98 percent of respondents in both the US and the UK either totally or somewhat agreed that the instructions were clear, indicating that the survey was well understood by the vast majority of participants.

For robustness, Figure A9 shows the main results excluding 40 participants in the US and 18 participants in the UK who responded "totally disagree" with the statement "The instructions were clear and easy to understand." The results remain robust to the exclusion of these respondents.

In addition, we analyzed answers to the follow-up open-ended question asking whether there was anything unclear or confusing about the study. Responses were coded using a combination of automated text matching and manual review. Specifically, we classified responses as indicating no concerns when respondents explicitly stated that nothing was unclear or confusing (e.g., "No," "None," "Nothing," "N/A," "All clear," "Everything was clear," "No issues," and similar expressions) or when they did not write anything. Using this approach, approximately 97 percent of respondents were coded as reporting no difficulties with the survey. We then manually inspected the remaining 3 percent of responses and found that many contained wording that was initially flagged by the automated procedure despite ultimately expressing that the survey was clear and easy to understand. The remaining comments primarily consisted of minor suggestions, requests for additional context, or observations about specific question wording rather than evidence of widespread confusion. Overall, these responses provide further reassurance that participants generally understood the survey instructions and questions.

Finally, respondents were asked whether there was anything else they would have liked to share with the researchers. Most people (90% in the US and 92% in the UK) did not share anything or wrote generic comments such as "No", "Thank you", or "No comments". Manual inspection of the remaining open-ended comments revealed no evidence that respondents systematically

misunderstood the survey. One of the most common suggestions concerned the difficulty of expressing nuanced views using binary or quantitative response.

A.6 Additional design considerations

A.6.1 List example provided to the participants

To ensure that respondents fully understood the task, they were first presented with an example list-experiment question and guided through the logic underlying a sample numerical response. Participants were reminded that they should report only the total number of statements that applied to them, rather than identifying which specific statements were true. This example is reported in Online Appendix D.

Before proceeding to the list-experiment questions, all respondents were given the opportunity to review the instructions once more. This additional step was intended to reinforce understanding of the task and reduce the likelihood of response errors.

A.6.2 Standard errors and weights

Differences involving list-experiment estimates, as done in Figures 1-2, are computed using bootstrap procedures. Bootstrap inference is appropriate because the double-list estimator is constructed as a combination of multiple sample means. Furthermore, the estimators being compared are often calculated from different or only partially overlapping samples. For example, direct-question estimates are based on all respondents who answered the direct question, including some who did not participate in the list experiment, whereas comparisons across key statements involve separate groups of respondents because each respondent was exposed to only one key statement. Because the sampling covariance between these estimators depends on the overlap in their estimation samples, standard post-estimation procedures such as *lincom* or conventional t-tests do not directly provide valid inference for their differences. The bootstrap accounts for the full estimation procedure, including the construction of the double-list estimator, the differing estimation samples, and the covariance arising from overlapping estimation samples.

The US and UK samples were recruited using quota sampling based on sex, age, and political affiliation. Because respondents were not selected through a probability sampling design, and because no survey weights or inclusion probabilities were available, we do not use survey-design estimators (*svy*). Instead, following common practice for quota-based online surveys, we account for the quota variables directly in the regression models. Specifically, Tables C2 and C3 include controls for sex, age group, and political affiliation, alongside other demographic and socioeconomic covariates. Reassuringly, the estimated list-experiment effects are very similar to those obtained without these controls.

In addition, the main estimates and standard errors reported in Table C1 do not change substantially when the Stata command *kict ls* (Tsai 2019) is used to perform least squares estimation specifically for a double list experiment.

A.6.3 Advantages of survey experiments and online surveys

As emphasized by Stantcheva (2023), surveys are not merely tools for data collection; they also enable researchers to generate controlled variation and exercise substantial control over the data-generation process. Moreover, although administrative data are invaluable for many research purposes, they cannot directly measure perceptions, beliefs, attitudes, knowledge, or reasoning. Likewise, while economists often emphasize revealed preferences, many important determinants of social, economic, and political outcomes are difficult to infer from observed behavior alone. Surveys therefore provide an important complement to administrative and other observational data, particularly in the study of transgender health and transgender-related policies, where underlying mechanisms can be difficult to identify using aggregate statistics alone.

Stantcheva (2023) also highlights several advantages of online surveys relative to in-person, telephone, and mail surveys. Online surveys allow respondents to participate at their convenience, reducing selection based on availability during working hours or willingness to answer the door or telephone. They may therefore facilitate participation among individuals balancing multiple responsibilities, such as caregivers. The accessibility of mobile technologies can also attract respondents who might otherwise be unwilling to complete traditional surveys. In addition, online surveys can reach populations that are often difficult to interview in person, including younger individuals, highly mobile populations, and residents of remote areas. Finally, survey platforms typically provide incentives for participation, helping to broaden the pool of potential respondents.

A common concern regarding online quota-based samples such as those obtained on Prolific because respondents are not selected through a probability-based sampling procedure, it is not possible to calculate conventional sampling errors or confidence intervals that are fully representative of the general population. In addition, such samples exclude individuals who do not use the internet. However, as noted by Haaland, Roth, and Wohlfart (2023), the practical differences between probability-based panels and quota-based online samples may be smaller than often assumed, given the substantial nonresponse rates typically observed in probability-based surveys. In the United Kingdom, for example, falling response rates and related quality concerns in the Labour Force Survey led the Office for National Statistics to temporarily suspend labour-market publications based on the survey in 2023 (Office for National Statistics, 2024). Furthermore, their review of the literature finds that online and offline populations tend to exhibit remarkably similar survey responses and experimental outcomes. It is also noteworthy that internet penetration in the UK and the US is very high: approximately 95% of the population used the internet in both countries in 2024 (World Bank, 2026). Taken together, these considerations

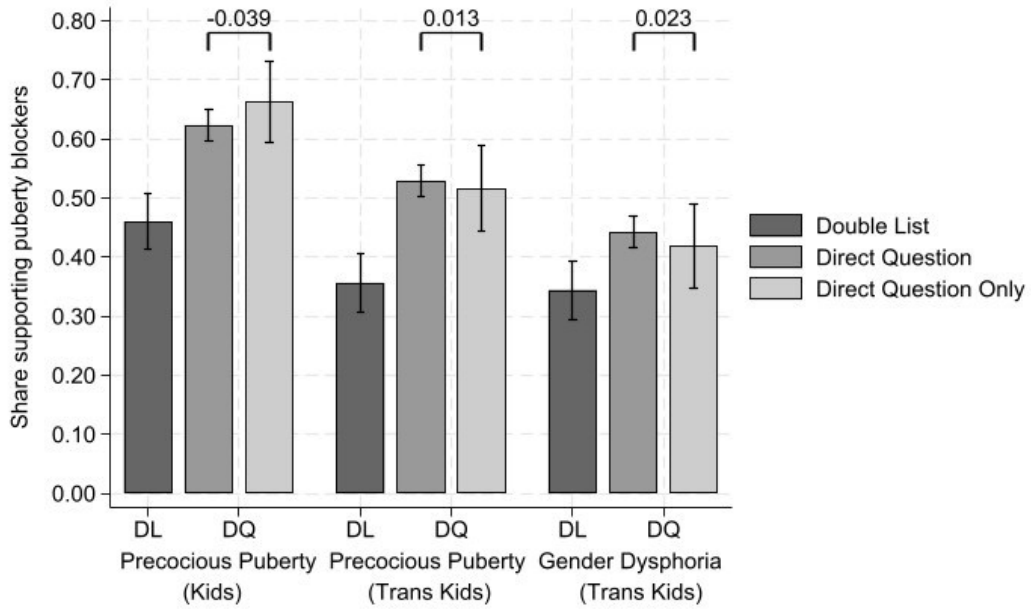
suggest that concerns about the representativeness of online samples may be less severe than is often presumed.

A.6.4 Randomized response technique

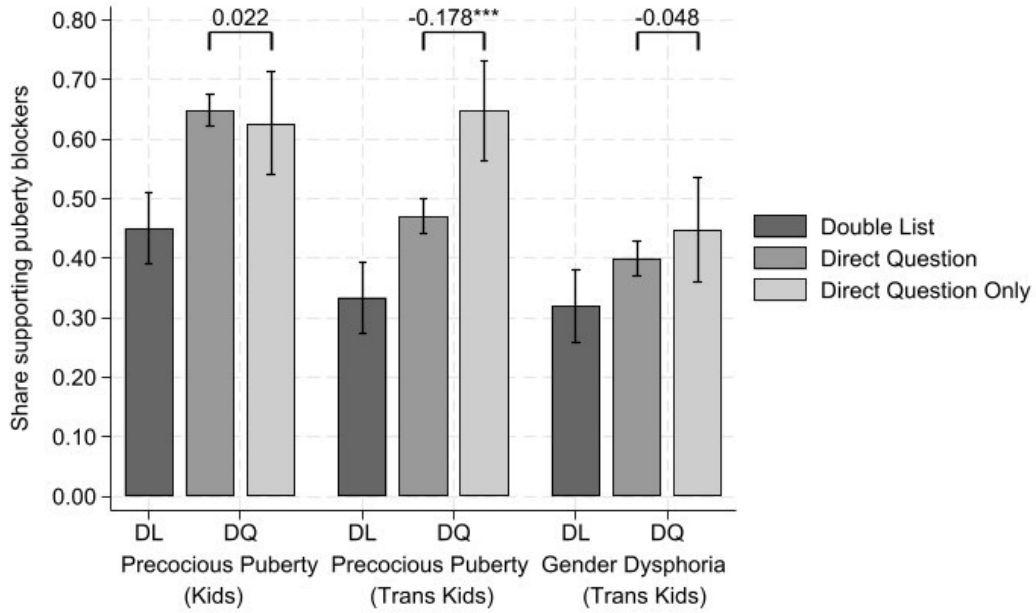
This study employs list experiments rather than the randomized response technique, in which respondents use a private randomization device, such as a coin flip, to determine whether they answer a sensitive or a non-sensitive question. This choice was motivated by three considerations, in addition to the advantages of list experiments discussed in Online Appendix A.4.1. First, the randomized response technique is generally more difficult to implement in online survey settings. Second, existing evidence suggests that respondents place greater trust in the anonymity provided by list experiments than in that provided by the randomized response technique (Coutts and Jann, 2011). Third, studies have shown that participants do not always comply with the randomization procedure required by the randomized response technique, potentially undermining the validity of the resulting estimates (John et al., 2018).

Figure A1: Comparing respondents who saw and did not see the list experiment

Panel A: US



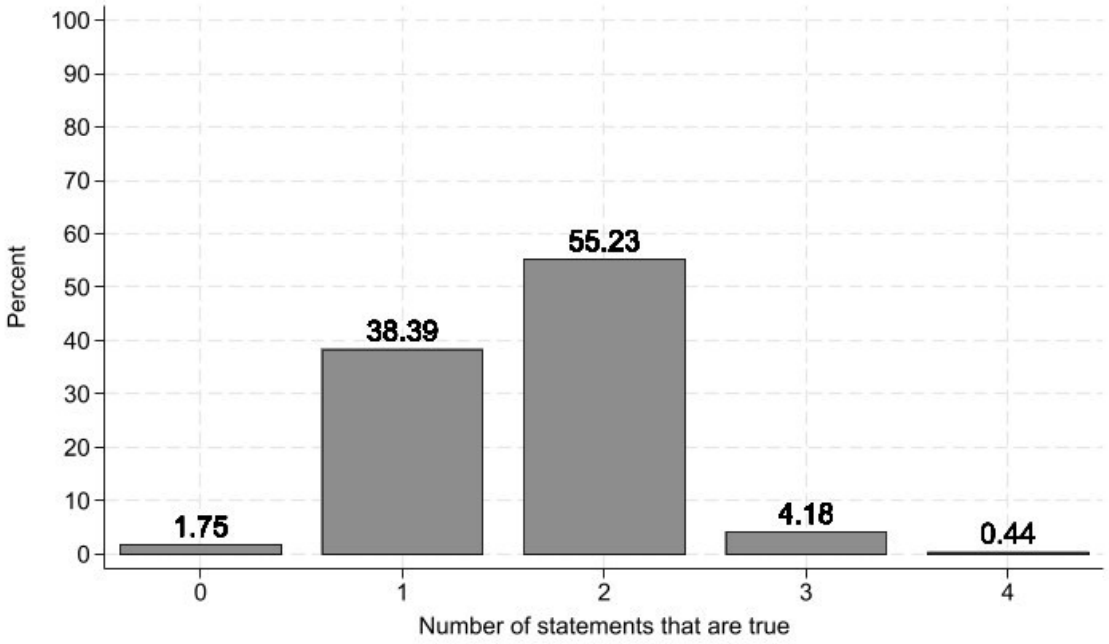
Panel B: UK



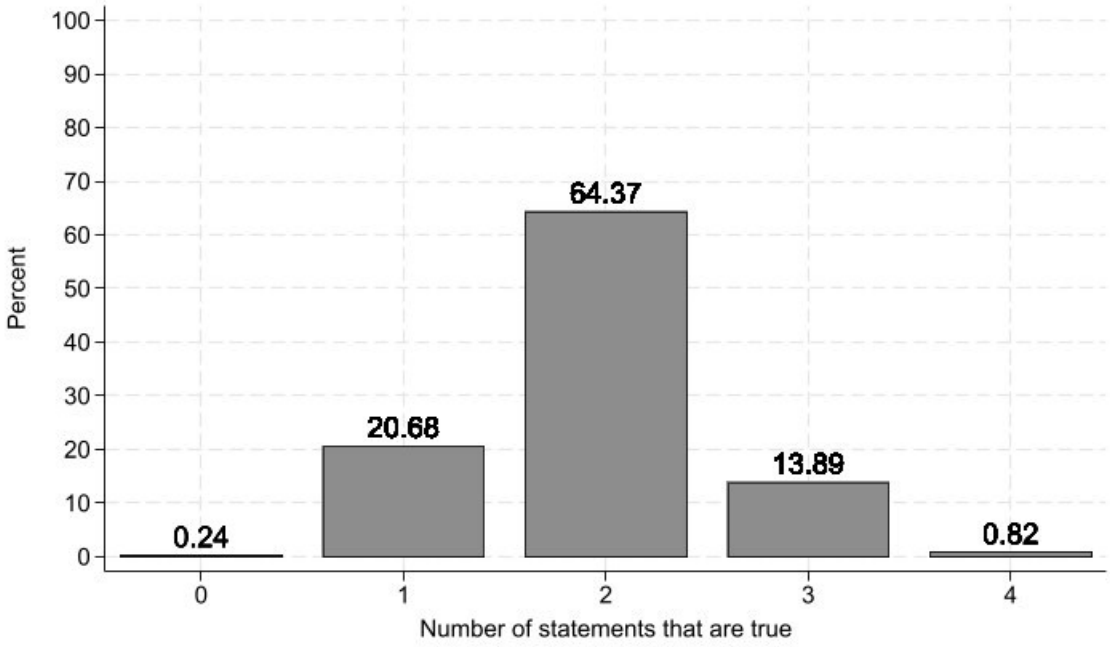
Notes: Unweighted statistics. The figure compares the list experiment estimate with direct-question responses among respondents who first completed the list experiment ("Direct") and respondents who answered only the direct question ("Direct Only"). The reported differences above the bars compare the two direct-question estimates. 95-percent confidence intervals are reported via the vertical range plots. Number of observations in Panel A: 1,337 (Bar 1); 1,298 (Bar 2); 181 (Bar 3) 1,407 (Bar 4); 1,355 (Bar 5); 182 (Bar 6); 1,385 (Bar 7); 1,355 (Bar 8); 186 (Bar 9). Number of observations in Panel B: 1,192 (Bar 1); 1,260 (Bar 2); 1,191 (Bar 3); 128 (Bar 4); 1,194 (Bar 5); 1,248 (Bar 6).

Figure A2: Distribution of responses to List Experiment - Without Key Statement - US

Panel A: List A



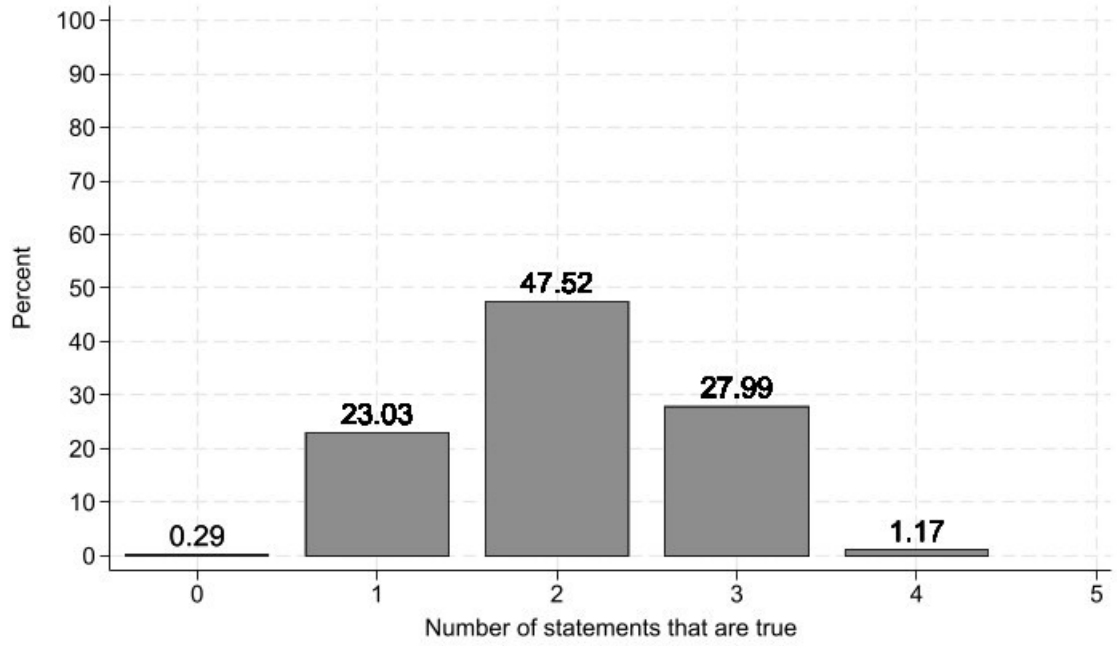
Panel B: List B



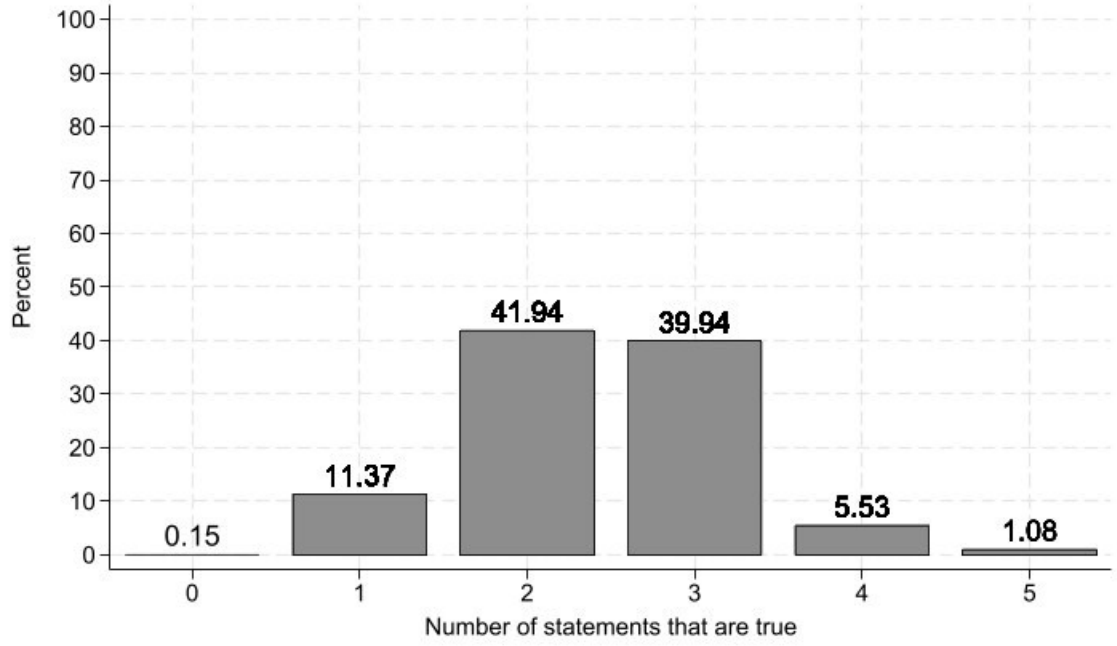
Notes: Unweighted statistics. Number of observations: 2,055 (Panel A), 2,074 (Panel B).

Figure A3: Distribution of responses to List Experiment - With Key Statement - US

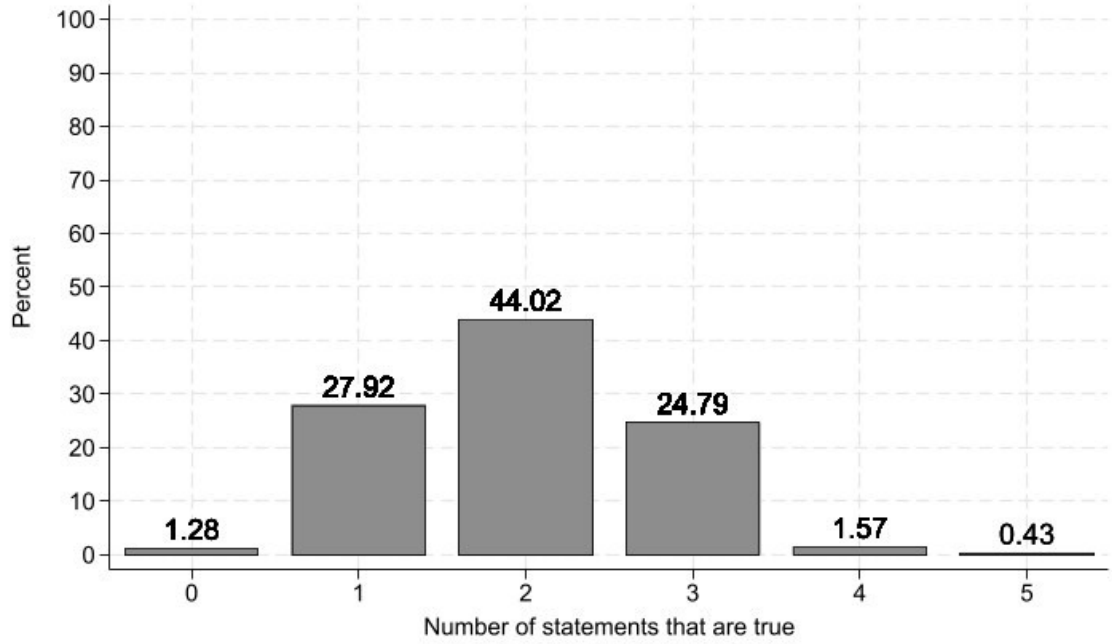
Panel A: Key statement 1 - List A



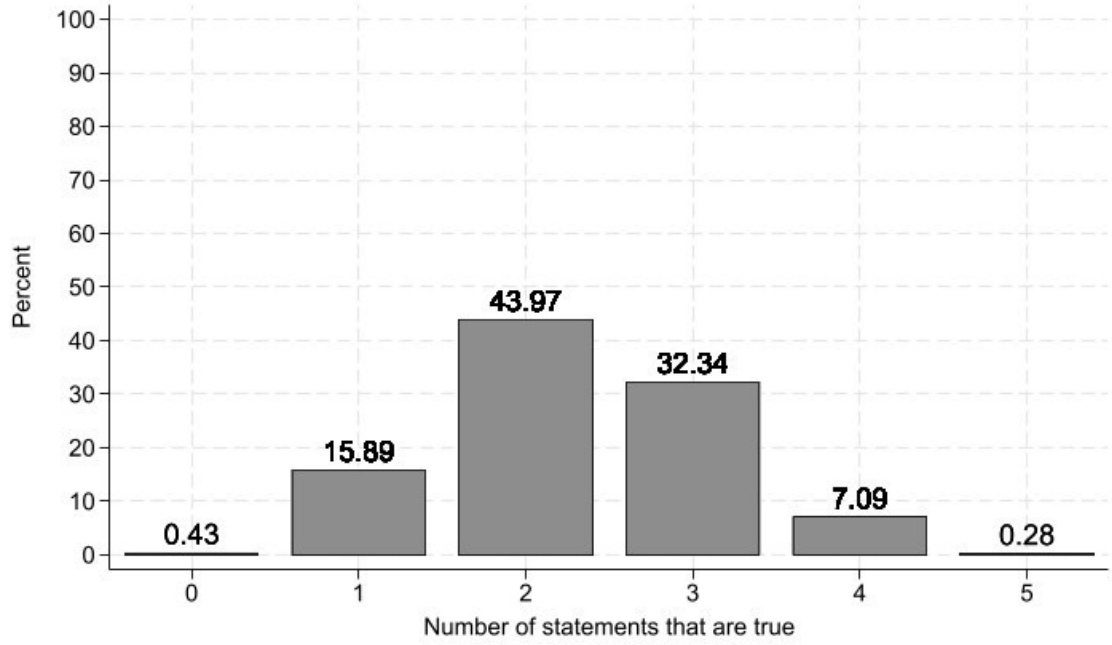
Panel B: Key statement 1 - List B



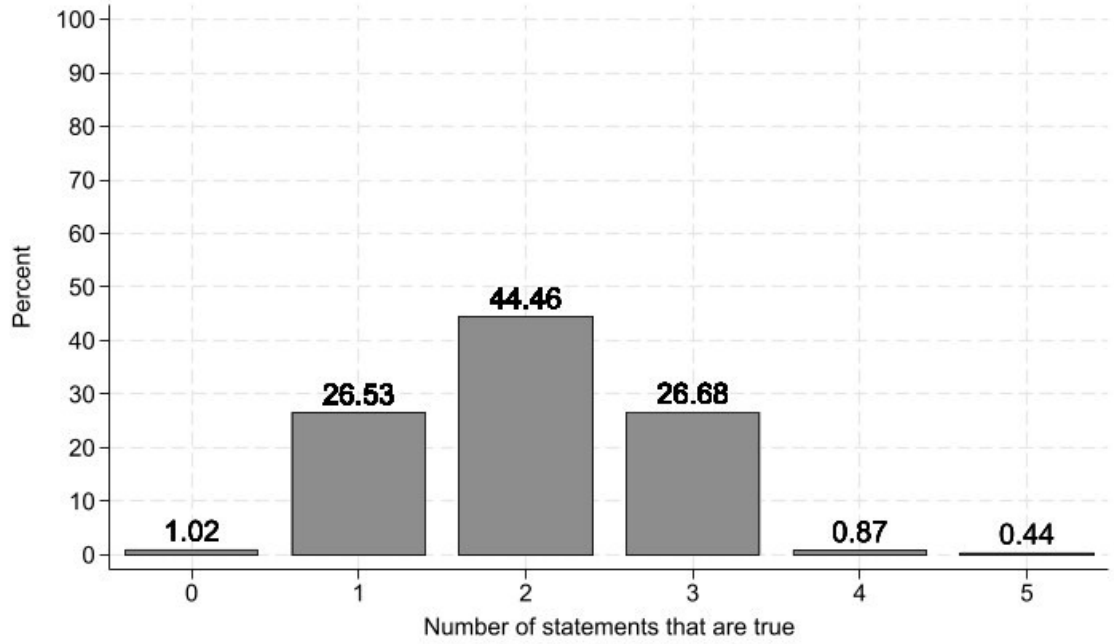
Panel C: Key statement 2 - List A



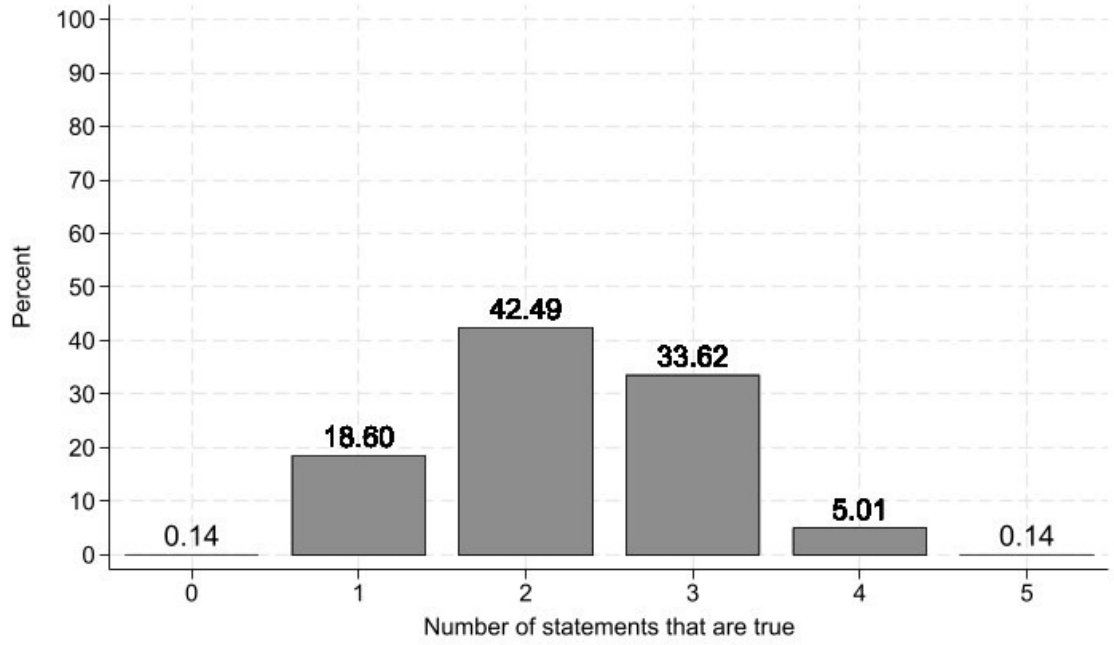
Panel D: Key statement 2 - List B



Panel E: Key statement 3 - List A



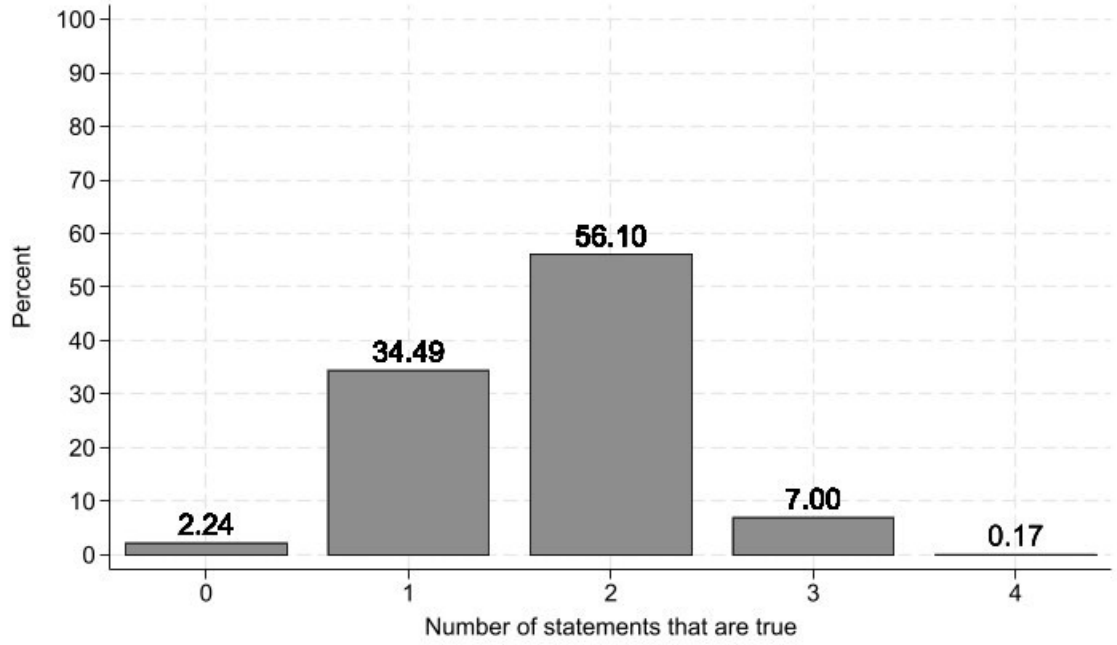
Panel F: Key statement 3 - List B



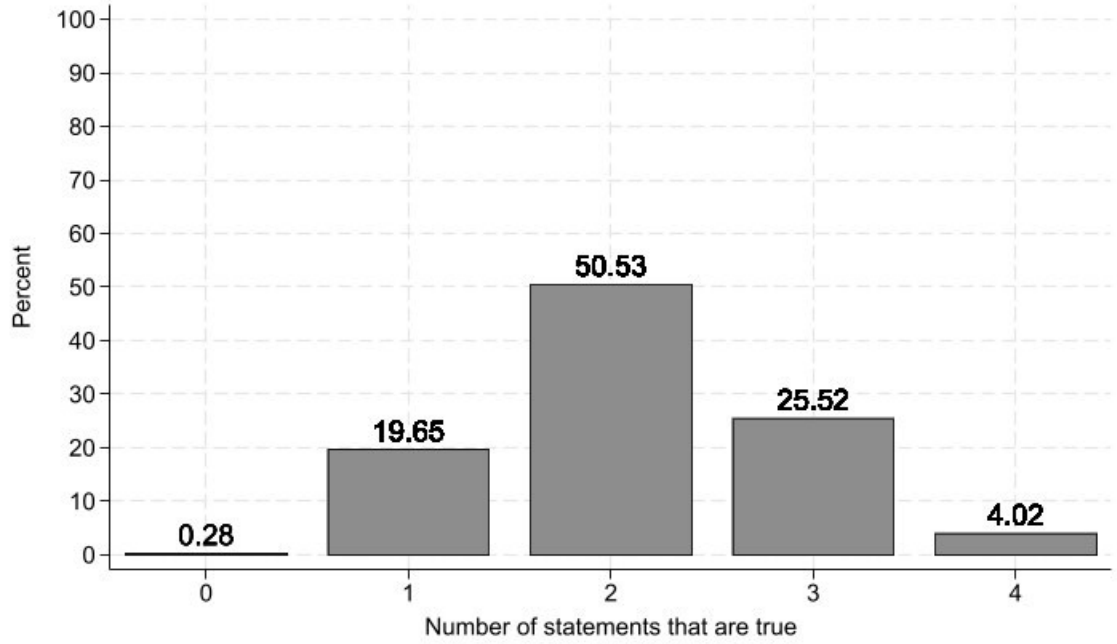
Notes: Unweighted statistics. Number of observations: 686 (Panel A), 651 (Panel B), 702 (Panel C), 705 (Panel D), 686 (Panel E), 699 (Panel F).

Figure A4: Distribution of responses to List Experiment - Without Key Statement - UK

Panel A: List A



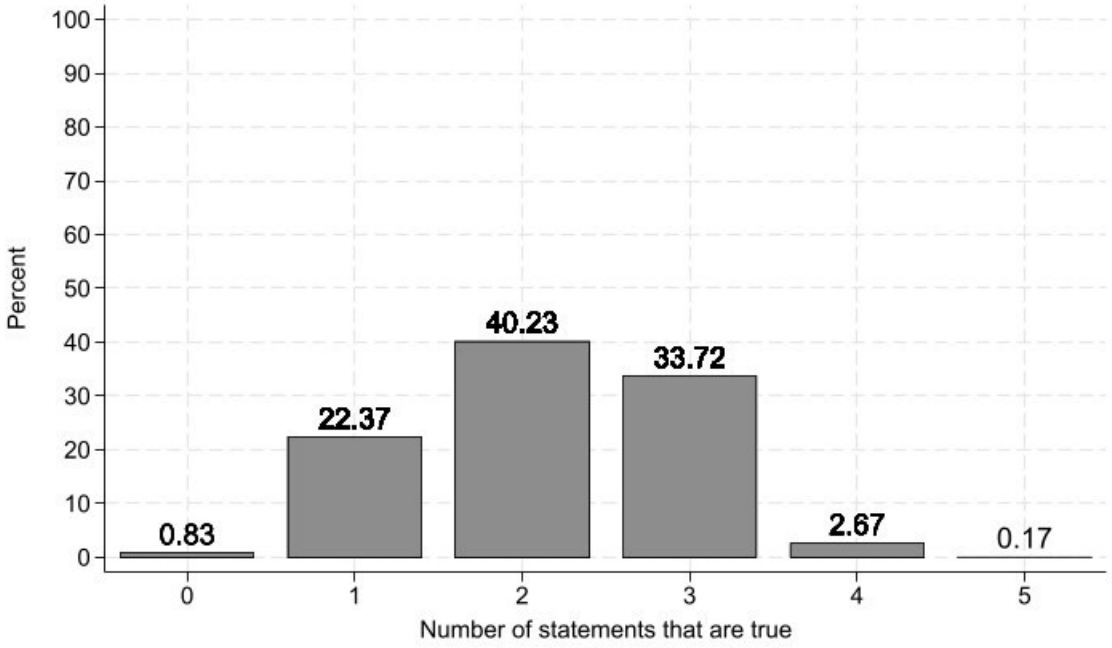
Panel B: List B



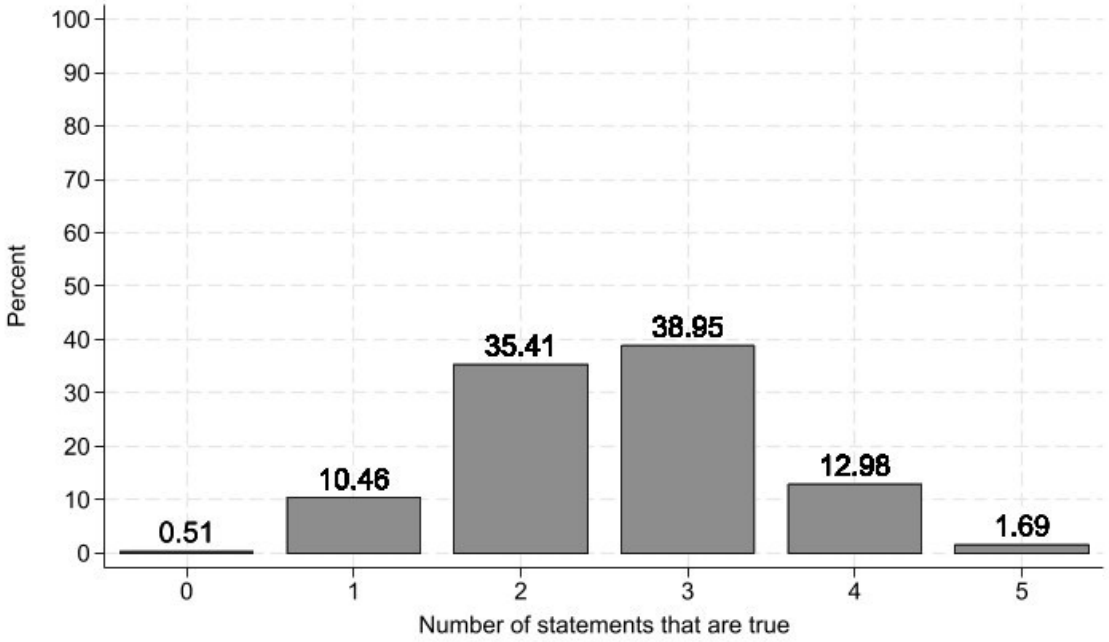
Notes: Unweighted statistics. Number of observations: 1,786 (Panel A), 1,791 (Panel B).

Figure A5: Distribution of responses to List Experiment - With Key Statement - UK

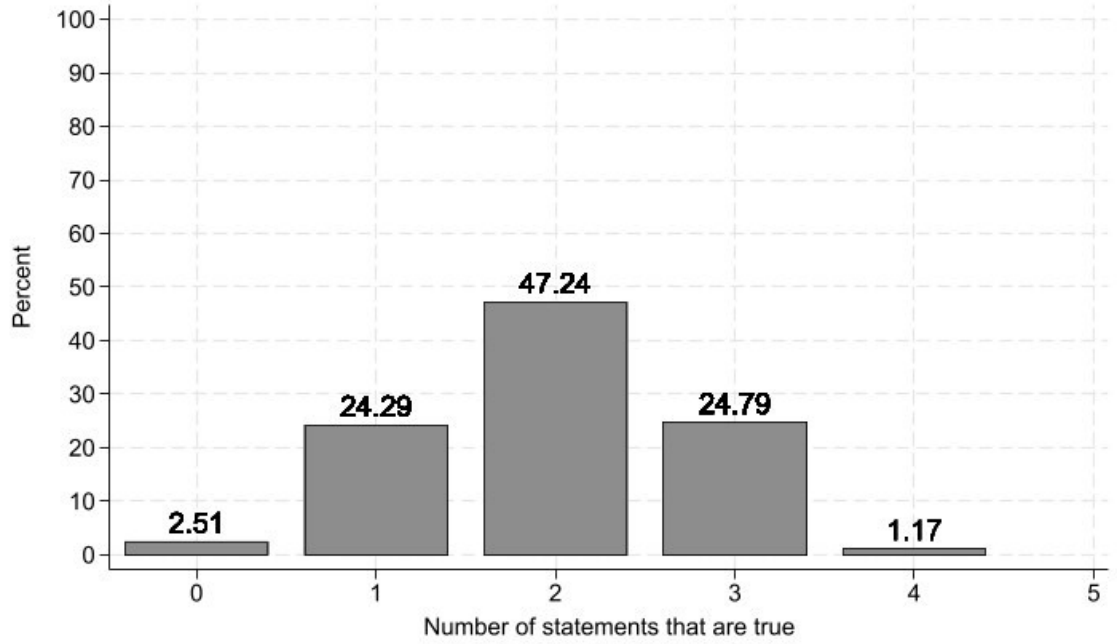
Panel A: Key statement 1 - List A



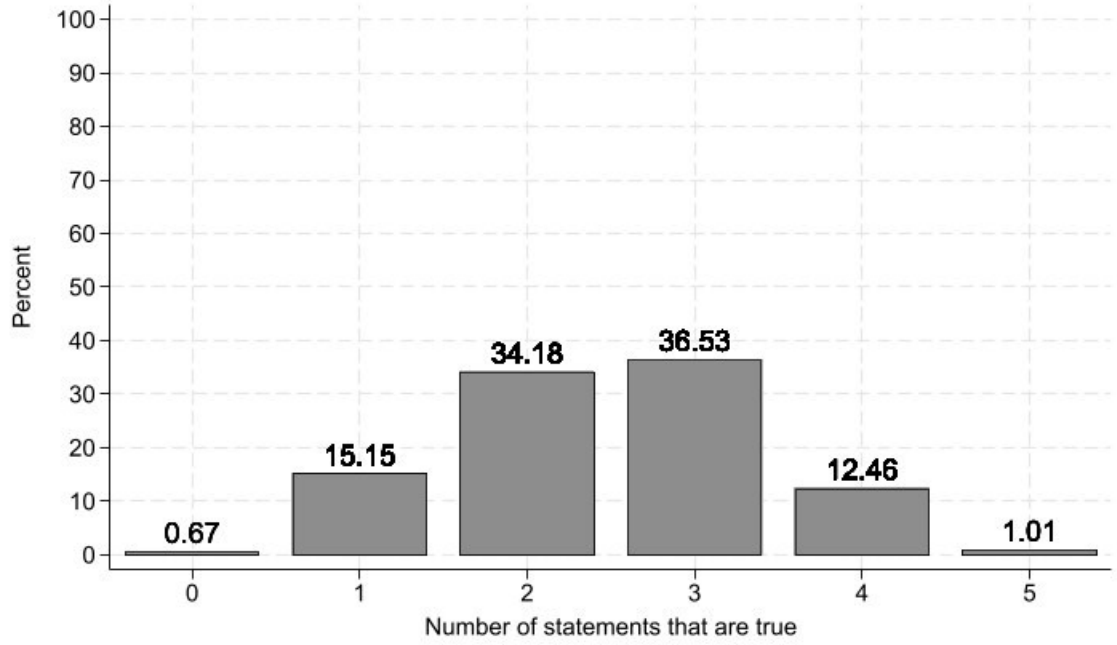
Panel B: Key statement 1 - List B



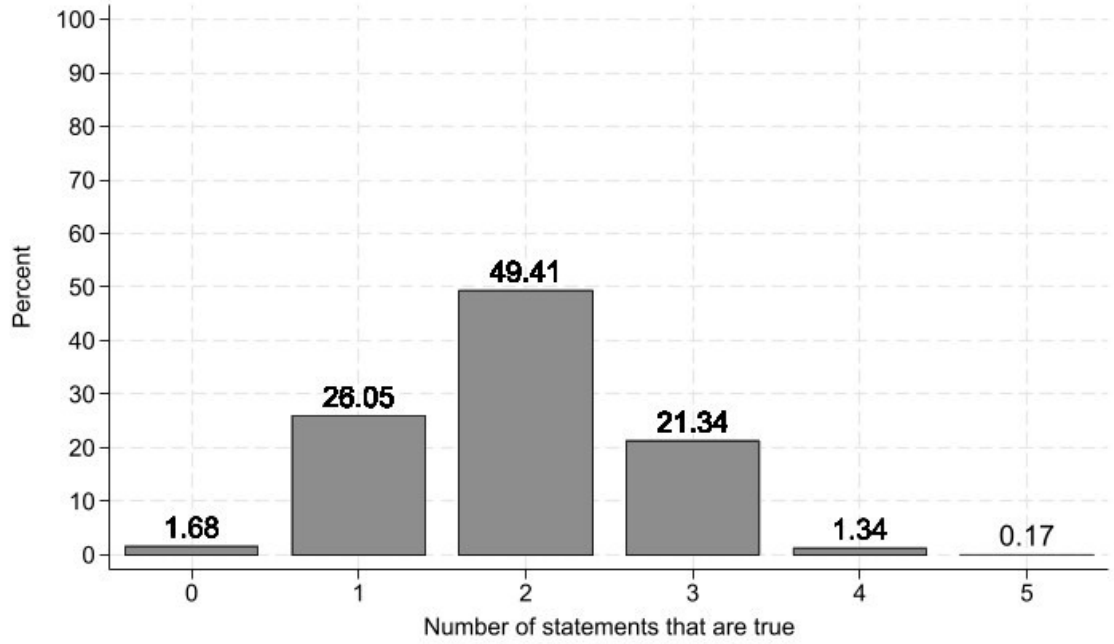
Panel C: Key statement 2 - List A



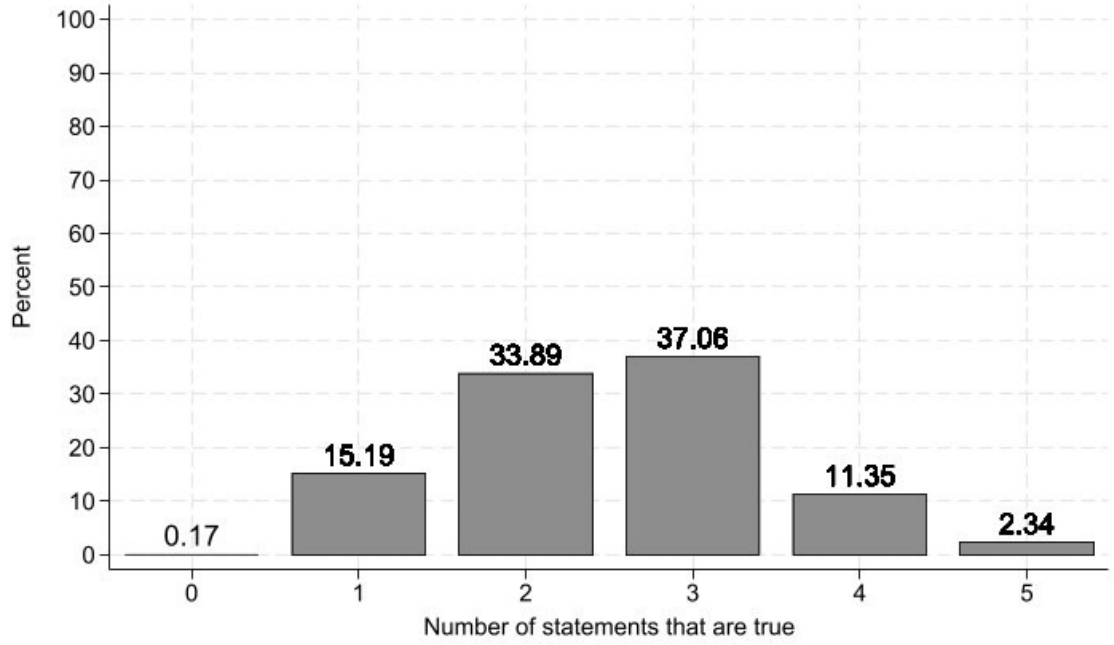
Panel D: Key statement 2 - List B



Panel E: Key statement 3 - List A



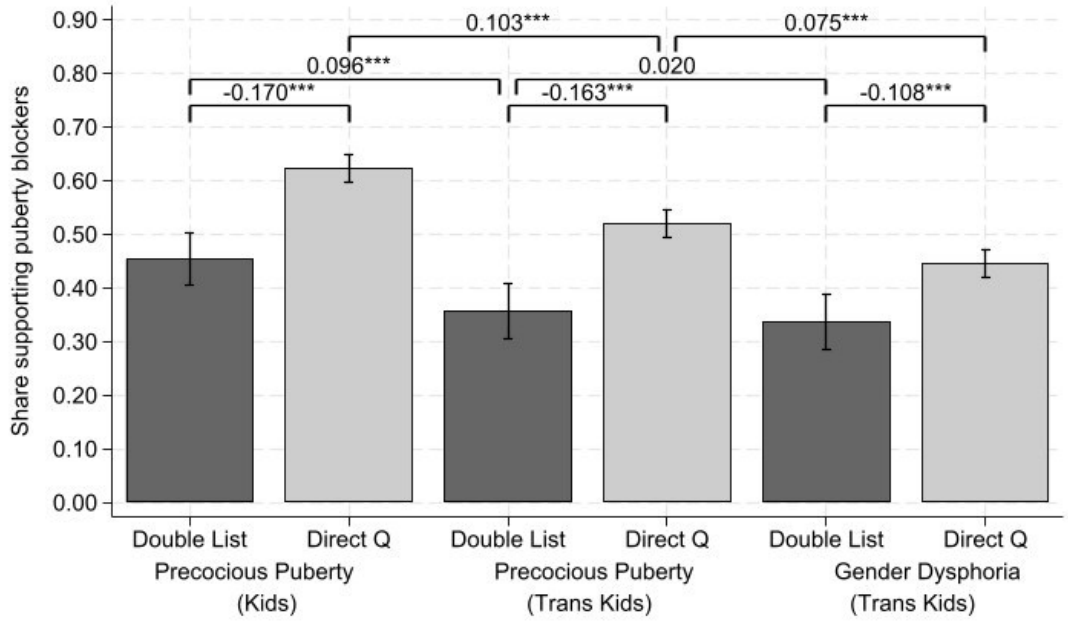
Panel F: Key statement 3 - List B



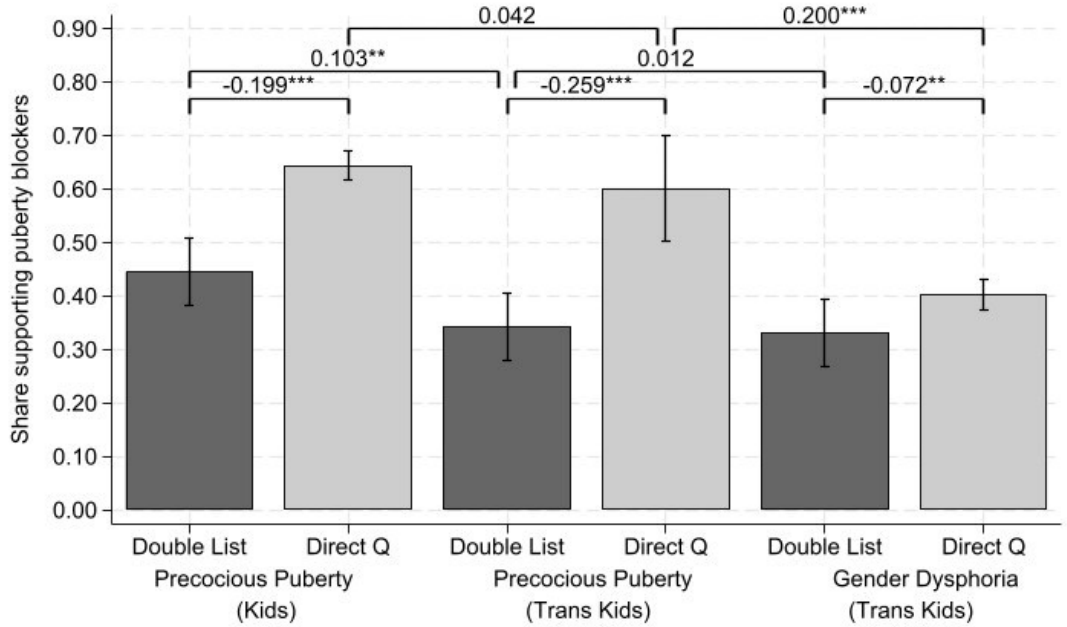
Notes: Unweighted statistics. Number of observations: 599 (Panel A), 593 (Panel B), 597 (Panel C), 594 (Panel D), 595 (Panel E), 599 (Panel F).

Figure A6: Exclude speeders and procrastinators

Panel A: US



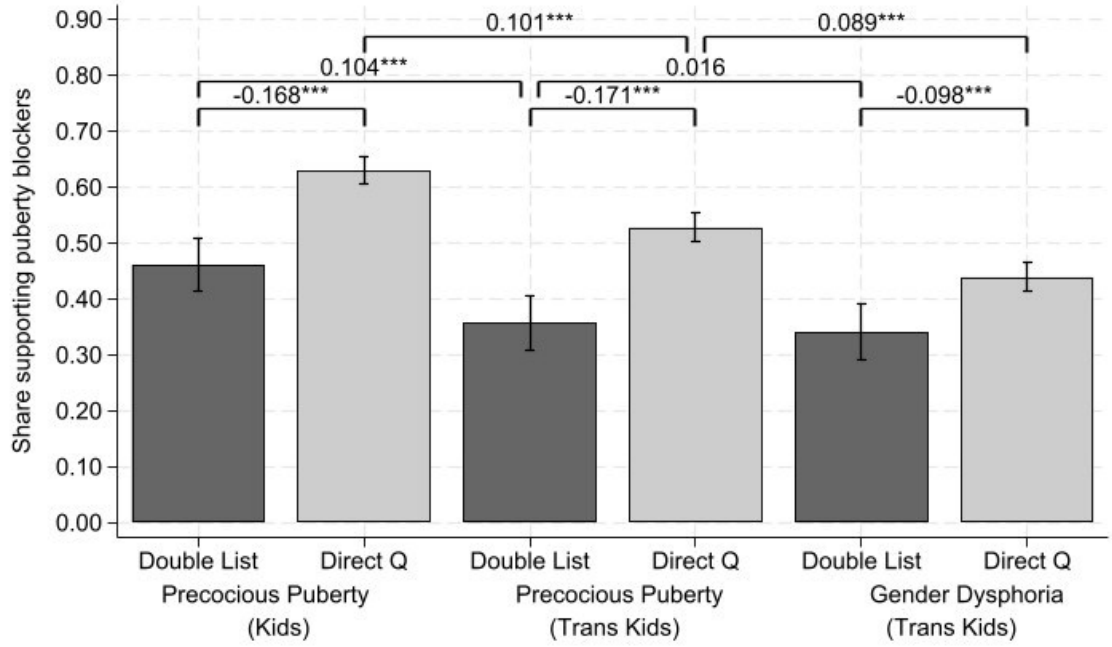
Panel B: UK



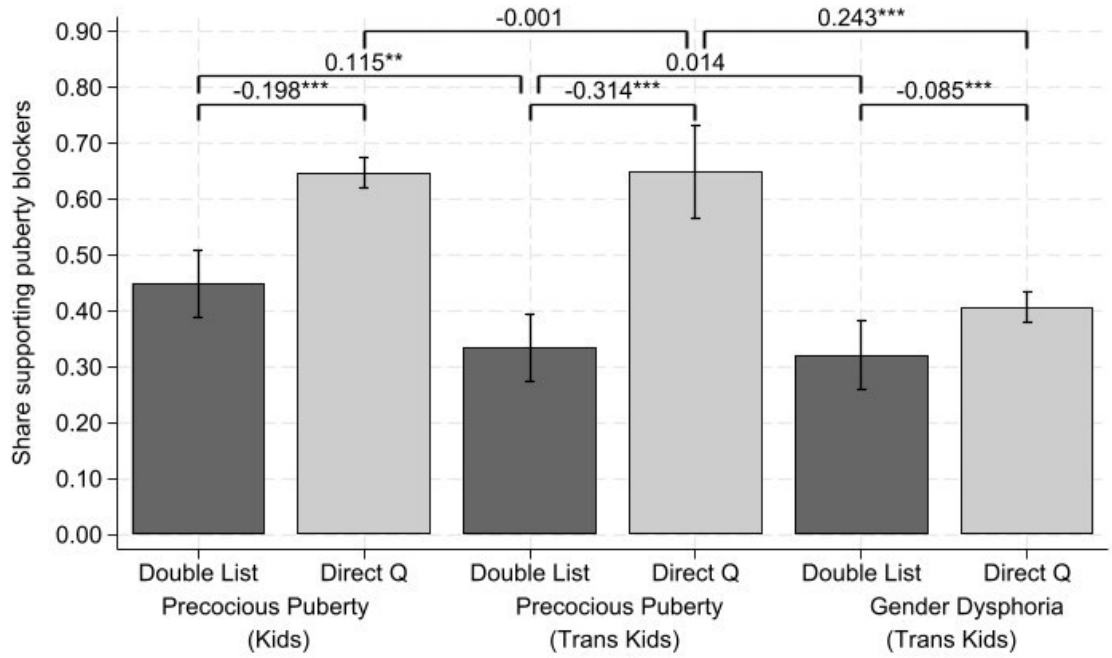
Notes: Unweighted statistics. This figure excludes 486 participants in the US and 423 participants in the UK in the top and bottom 5 percent of the completion-time distribution. See also notes in Figures 1-2.

Figure A7: Exclude respondents who failed more than two attention checks

Panel A: US



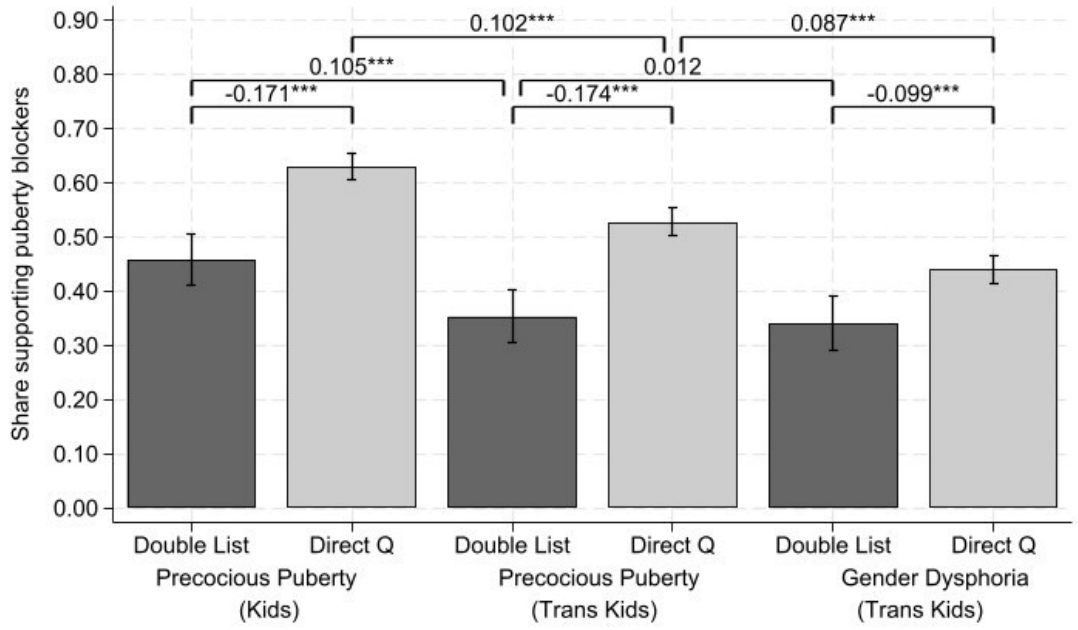
Panel B: UK



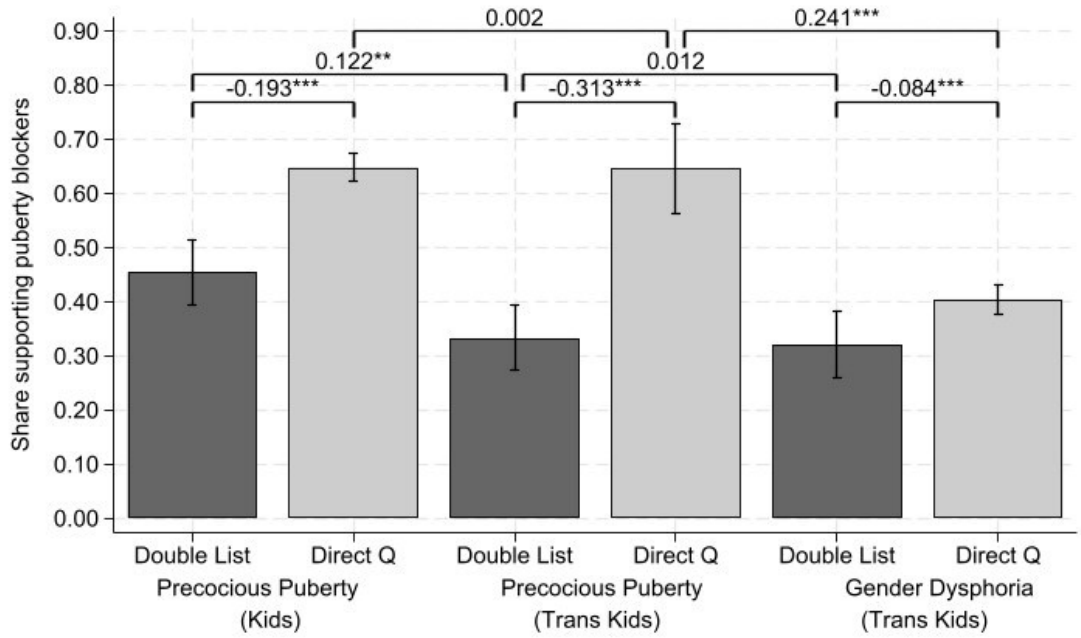
Notes: Unweighted statistics. This figure excludes 6 participants in the US and 3 participants in the UK who failed two or more attention checks. See also notes in Figures 1-2.

Figure A8: Exclude respondents who scored less than 0.8 in the CAPTCHA verification

Panel A: US



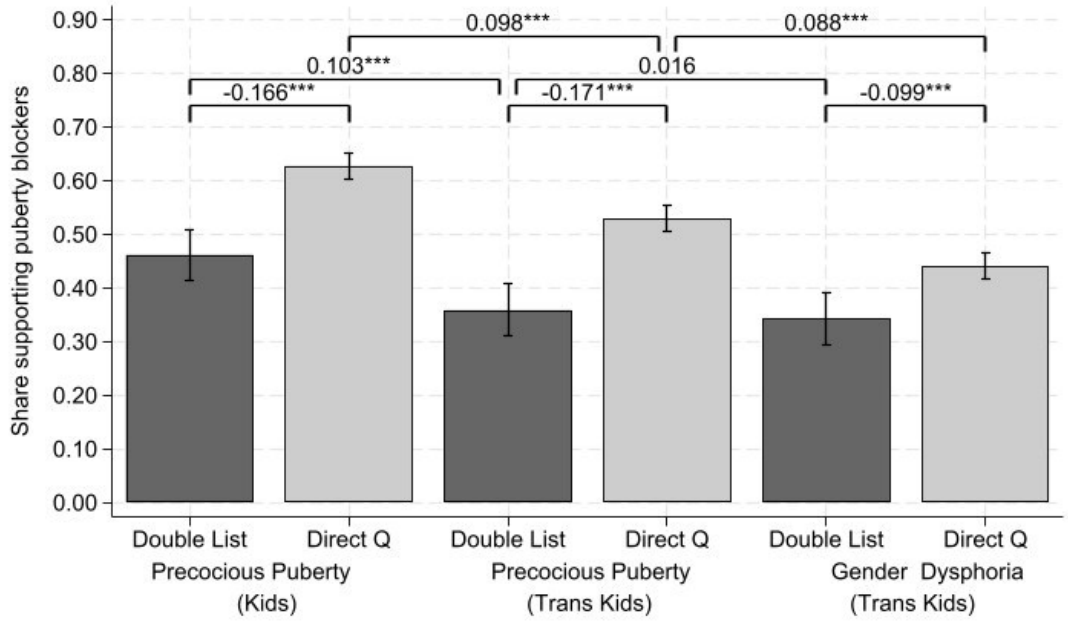
Panel B: UK



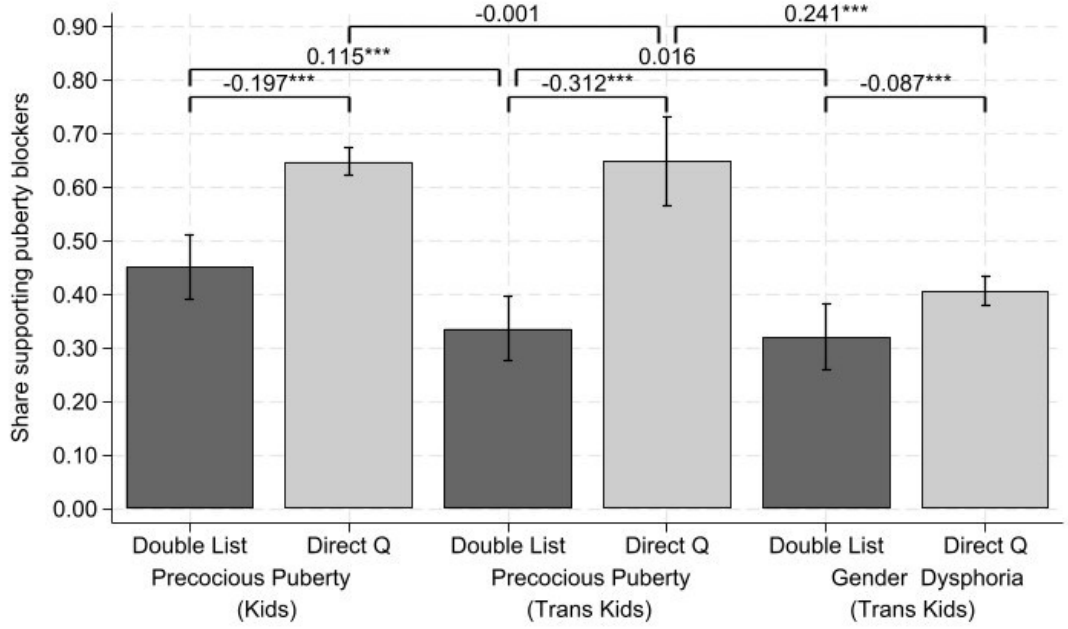
Notes: Unweighted statistics. This figure excludes 75 participants in the US and 47 participants in the UK who scored less than 0.5 in the CAPTCHA verification or whose CAPTCHA score was missing. See also notes in Figures 1-2.

Figure A9: Exclude respondents who found the survey instructions unclear

Panel A: US



Panel B: UK



Notes: Unweighted statistics. This figure excludes 40 participants in the US and 18 participants in the UK who responded "totally disagree" with the statement "The instructions were clear and easy to understand." See also notes in Figures 1-2.

Table A1: Balance Table - US

Variable	List A with key statement (List B without key statement)	List A without key statement (List B with key statement)	Difference
Age			
Mean	46.22	46.30	-0.080
between 18-34	0.293	0.295	-0.002
between 35-49	0.252	0.250	0.002
between 50-64	0.320	0.310	0.010
65 or older	0.136	0.146	-0.010
Female (sex at birth)	0.511	0.523	-0.012
Race			
White only	0.783	0.768	0.016
Black or African American only	0.105	0.125	-0.019**
Asian or Native Hawaiian or Pacific Islander only	0.051	0.051	0.001
Never married	0.363	0.372	-0.009
Higher education	0.531	0.533	-0.002
Employed	0.690	0.703	-0.014
Household income: less than $60,000 (US)	0.425	0.430	-0.005
Political Party Affiliation			
Democrat	0.317	0.328	-0.011
Republican	0.294	0.276	0.018
Independent	0.389	0.396	-0.007
Religiosity - Not at all important	0.334	0.323	0.011
Urbanicity			
Rural area	0.162	0.162	0.001
Small city or town	0.262	0.271	-0.010
Suburb near a large city	0.378	0.364	0.014
Large city	0.198	0.203	-0.005
Region			
Northeast	0.177	0.179	-0.001
Midwest	0.208	0.203	0.005
South	0.426	0.438	-0.011
West	0.189	0.180	0.008
Total number of participants	2,751	2,074	

Notes: Unweighted statistics. See also notes in Table 2. *p < 0.10; **p < 0.05; ***p < 0.01.

Table A2: Balance Table - UK

Variable	List A with key statement (List B without key statement)	List A without key statement (List B with key statement)	Difference
Age			
Mean	46.48	46.50	-0.025
between 18-34	0.282	0.271	0.011
between 35-49	0.247	0.261	-0.015
between 50-64	0.339	0.350	-0.011
65 or older	0.133	0.118	0.015
Female (sex at birth)	0.513	0.543	-0.030*
Race			
White only	0.866	0.855	0.011
Black or Caribbean	0.041	0.050	-0.009
Asian or Arab	0.060	0.056	0.004
Never married	0.407	0.421	-0.014
Higher education	0.626	0.618	0.008
Employed	0.712	0.733	-0.021
Household income: less than £36,000	0.375	0.386	-0.011
Political Party Affiliation			
Labour/Green/LibDem	0.600	0.594	0.007
Tory/Reform UK	0.317	0.311	0.006
Other	0.083	0.095	-0.013
Religiosity - Not at all important	0.577	0.560	0.017
Urbanicity			
Rural area	0.186	0.190	-0.004
Small city or town	0.439	0.409	0.029*
Suburb near a large city	0.190	0.207	-0.017
Large city	0.185	0.193	-0.008
Region			
England	0.856	0.866	-0.010
Scotland	0.080	0.071	0.009
Northern Ireland	0.014	0.016	-0.001
Wales	0.049	0.047	0.002
Total number of participants	2,392	1,791	

Notes: Unweighted statistics. See also notes in Table 2. *p < 0.10; **p < 0.05; ***p < 0.01.

Table A3: No design effect - Blair and Imai (2012) test

		List A				List B			
		p-value		p-value (No GMS)		p-value		p-value (No GMS)	
Country	Key statement	Pr(R ,S=0)	Pr(R ,S=1)	Pr(R ,S=0)	Pr(R ,S=1)	Pr(R ,S=0)	Pr(R ,S=1)	Pr(R ,S=0)	Pr(R ,S=1)
US	KS1	N.E.	N.E.	N.E.	N.E.	0.252	1.000	0.644	0.906
US	KS2	0.325	1.000	0.737	0.899	1.000	0.159	0.884	0.504
US	KS3	0.320	1.000	0.731	0.901	1.000	1.000	0.888	0.906
UK	KS1	0.159	1.000	0.505	0.901	1.000	0.041	0.871	0.196
UK	KS2	N.E.	N.E.	N.E.	N.E.	1.000	0.205	0.868	0.570
UK	KS3	1.000	1.000	0.887	0.896	1.000	1.000	0.862	0.889

Notes: Unweighted statistics. The table reports p-values from the Blair and Imai (2012) design-effect test implemented using *kict deff*. The no-design-effect assumption requires respondents' answers to the non-key items to be unaffected by whether the key item is included in the list. The first set of columns reports p-values for tests of whether any estimated joint probabilities Pr(R=r,S=0) or Pr(R=r,S=1)are negative, where R denotes the number of non-key items endorsed and S indicates endorsement of the key item. Negative estimated probabilities would be inconsistent with the no-design-effect assumption. The null hypothesis is that none of the relevant joint probabilities is smaller than zero; therefore, large p-values indicate greater support for the identifying assumption. Bonferroni-adjusted p-values are reported both using the Generalized Moment Selection (GMS) procedure, which increases power by excluding probabilities that are clearly positive, and without the GMS adjustment ("No GMS"). N.E. (Not Estimable) indicates that the test could not be computed because no respondent reported the maximum possible count. This outcome is consistent with the absence of ceiling effects. As a robustness check, assigning a value of five to one randomly selected respondent who reported four allowed the test to be estimated and did not alter the substantive conclusion, as the null hypothesis continued to not be rejected.

Table A3: Feedback - Everything was clear and easy to understand

Panel A: US

	Freq.	Percent	Cum.
Totally Disagree	40	0.83	0.83
Somewhat Disagree	16	0.33	1.16
Neither Agree nor Disagree	26	0.54	1.70
Somewhat Agree	387	8.02	9.72
Totally Agree	4,356	90.28	100.00
Total	4,825	100.00	

Panel B: UK

	Freq.	Percent	Cum.
Totally Disagree	18	0.43	0.43
Somewhat Disagree	25	0.60	1.03
Neither Agree nor Disagree	19	0.45	1.48
Somewhat Agree	472	11.28	12.77
Totally Agree	3,649	87.23	100.00
Total	4,183	100.00	

Notes: Unweighted statistics.

References used in Appendix A

Office for National Statistics. 2024. "Impact of Reweighting on Labour Force Survey Key Indicators: 2024." Office for National Statistics, February 5, 2024. https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/articles/impactofreweightingonlabourforcesurveykeyindicators/2024

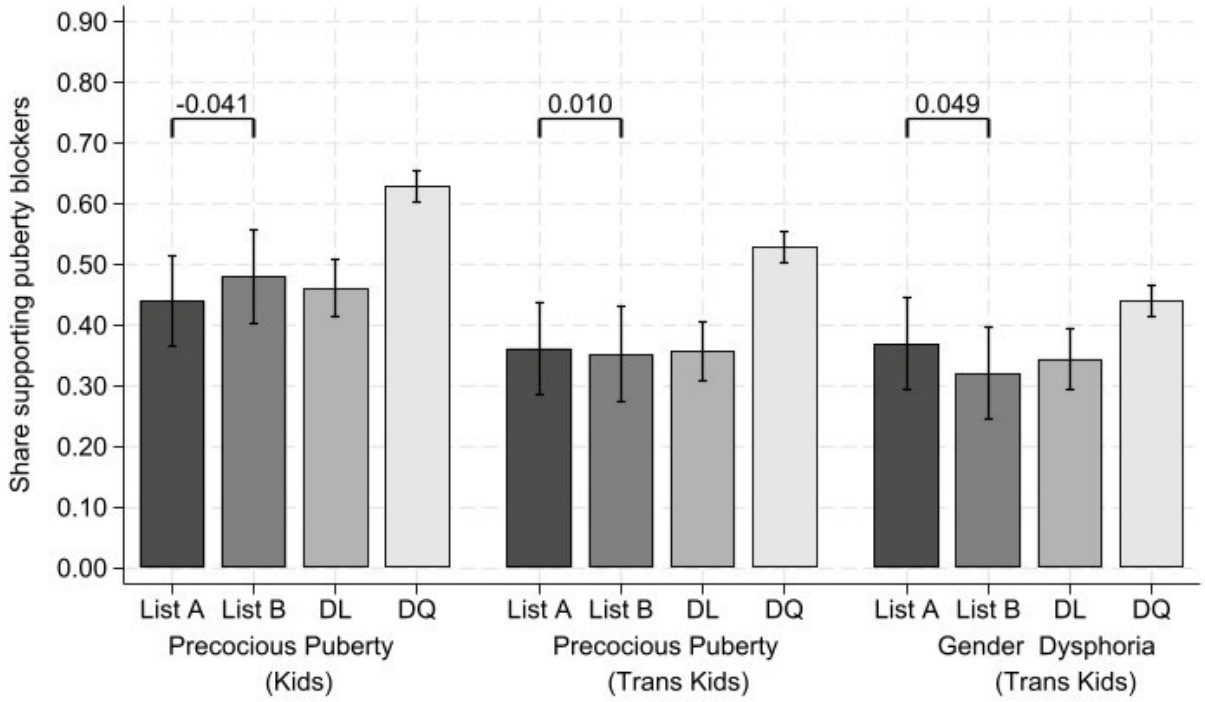
World Bank. 2026. "Individuals Using the Internet (% of Population)." *World Development Indicators*. https://data.worldbank.org/indicator/IT.NET.USER.ZS

Appendix B. Data Cleaning

To Be Added

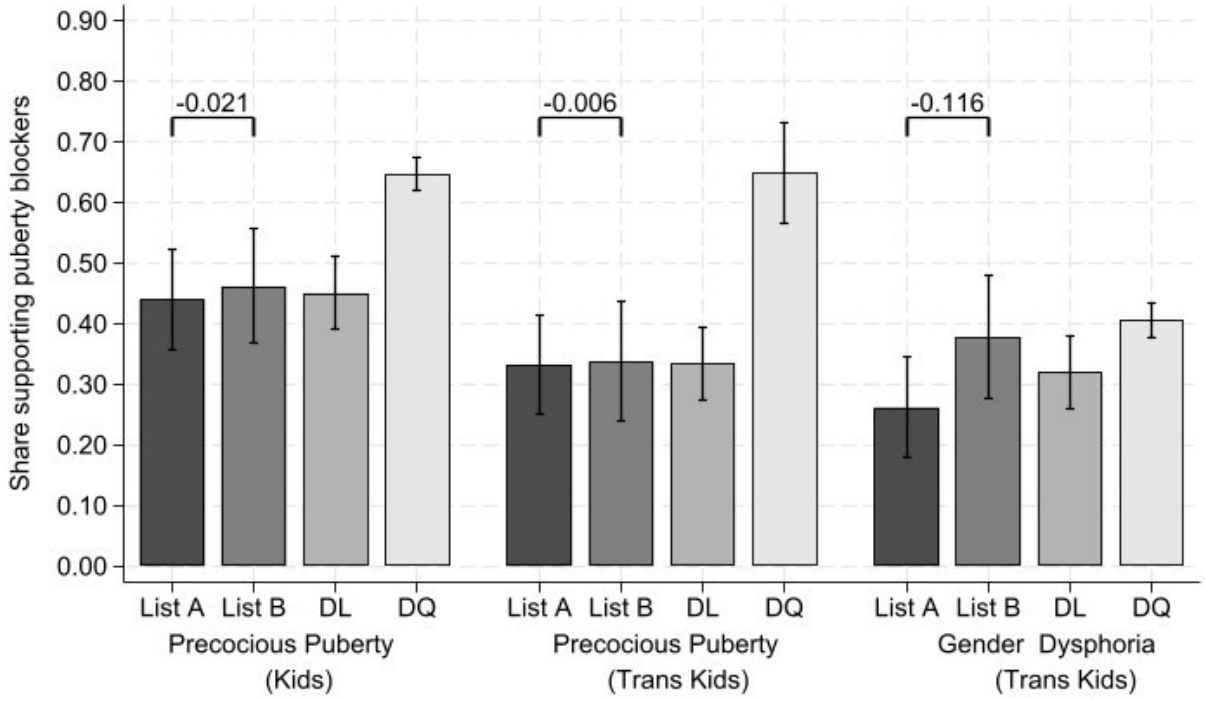
Appendix C. Additional Figures and Tables

Figure C1: Main list experiments including List A and List B - US



Notes: Unweighted statistics. In addition to replicating the double list ('DL') and direct question ('DQ') estimates reported in Figure 1, this figure reports the list estimates for Lists A and B separately for the three key statements. Number of observations: 1,337 (Bar 1-3); 1,479 (Bar 4); 1,407 (Bar 5-7); 1,537 (Bar 8); 1,385 (Bar 9-11); 1,521 (Bar 12).

Figure C2: Main list experiments including List A and List B - UK



Notes: Unweighted statistics. In addition to replicating the double list ('DL') and direct question ('DQ') estimates reported in Figure 2, this figure reports the list estimates for Lists A and B separately for the three key statements. Number of observations: 1,192 (Bar 1-3); 1,260 (Bar 4); 1,191 (Bar 5-7); 128 (Bar 8); 1,194 (Bar 9-11); 1,248 (Bar 12).

Table C1: Difference-in-means comparisons and robustness checks

	List A	List B	Double list	Direct question	(1)-(2)	(3)-(4)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: United States						
Precocious Puberty (Kids)	0.440	0.481	0.461	0.629	-0.041	-0.168***
	(0.038)	(0.040)	(0.024)	(0.013)		
Precocious Puberty (Trans Kids)	0.362	0.352	0.357	0.528	0.010	-0.171***
	(0.039)	(0.040)	(0.025)	(0.013)		
Gender Dysphoria (Trans Kids)	0.369	0.320	0.345	0.440	0.049	-0.095***
	(0.039)	(0.039)	(0.025)	(0.013)		
Panel B: United Kingdom						
Precocious Puberty (Kids)	0.440	0.462	0.451	0.647	-0.021	-0.196***
	(0.042)	(0.048)	(0.031)	(0.013)		
Precocious Puberty (Trans Kids)	0.332	0.337	0.335	0.648	-0.006	-0.314***
	(0.042)	(0.051)	(0.030)	(0.042)		
Gender Dysphoria (Trans Kids)	0.262	0.378	0.320	0.405	-0.116	-0.086***
	(0.042)	(0.052)	(0.031)	(0.014)		

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Bootstrapped standard errors in parentheses. See also notes in Figures 1-2.

Table C2: Main list experiments including List A and List B - Multivariate analysis - US

	List A				List B			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Precocious Puberty (Kids)								
Subject saw list with key statement	0.440***	0.425***	0.420***	0.404***	0.481***	0.502***	0.507***	0.567***
	(0.038)	(0.038)	(0.037)	(0.044)	(0.040)	(0.041)	(0.041)	(0.048)
Observations	1,337	1,337	1,337	1,298	1,337	1,337	1,337	1,298
R^2	0.092	0.142	0.175	0.194	0.101	0.140	0.171	0.219
Estimated bias	18.9	20.4	20.9	22.6	14.8	12.7	12.2	6.2
Panel B: Precocious Puberty (Trans Kids)								
Subject saw list with key statement	0.362***	0.375***	0.380***	0.391***	0.352***	0.341***	0.351***	0.348***
	(0.039)	(0.039)	(0.039)	(0.039)	(0.040)	(0.041)	(0.040)	(0.040)
Observations	1,407	1,407	1,407	1,355	1,407	1,407	1,407	1,355
R^2	0.059	0.115	0.142	0.171	0.052	0.105	0.151	0.183
Estimated bias	16.6	15.3	14.6	13.7	17.6	18.7	17.7	18.0
Panel C: Gender Dysphoria (Trans Kids)								
Subject saw list with key statement	0.369***	0.355***	0.356***	0.342***	0.320***	0.320***	0.317***	0.319***
	(0.039)	(0.039)	(0.039)	(0.039)	(0.039)	(0.040)	(0.040)	(0.040)
Observations	1,385	1,385	1,384	1,334	1,385	1,385	1,384	1,334
R^2	0.063	0.124	0.148	0.172	0.047	0.088	0.117	0.139
Estimated bias	7.1	8.5	8.4	9.8	12.0	12.0	12.3	12.1
Controls for:								
State fixed effects		✓	✓	✓		✓	✓	✓
Demographic controls		✓	✓	✓		✓	✓	✓
Socio-economic factors			✓	✓			✓	✓
Additional controls and beliefs				✓				✓

Notes: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. OLS estimates. Robust standard errors in parentheses. "Estimated bias" reports the differences (in percentage points) between the estimated percentage of participants who agreed to the key statement in each column and the estimate obtained from the corresponding direct question. *Demographic controls* include subject's age, sex at birth, race and ethnicity, and sexual orientation. *Socio-economic factors and beliefs* include subject's education level, employment status, income (1 person did not record their income), religiosity, and political affiliation. *Additional controls* include whether subject has a child, whether subject knows someone transgender, whether subject believes that more than 3% of individuals are transgender, urbanicity, marital status, day and week the survey was taken, and beliefs about general level of support for puberty blockers for precocious puberty (Panels A-B) or gender dysphoria (Panel C). See also notes in Figure 1.

Table C3: Main list experiments including List A and List B - Multivariate analysis - UK

	List A				List B			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Precocious Puberty (Kids)								
Subject saw list with key statement	0.440***	0.434***	0.435***	0.479***	0.462***	0.470***	0.462***	0.515***
	(0.042)	(0.042)	(0.042)	(0.052)	(0.048)	(0.048)	(0.048)	(0.060)
Observations	1,192	1,192	1,192	1,134	1,192	1,192	1,192	1,134
R^2	0.083	0.114	0.118	0.139	0.072	0.089	0.121	0.146
Estimated bias	20.7	21.3	21.2	16.8	18.5	17.7	18.5	13.2
Panel B: Precocious Puberty (Trans Kids)								
Subject saw list with key statement	0.332***	0.344***	0.339***	0.339***	0.337***	0.336***	0.337***	0.342***
	(0.042)	(0.041)	(0.041)	(0.042)	(0.051)	(0.051)	(0.049)	(0.050)
Observations	1,191	1,191	1,191	1,117	1,191	1,191	1,191	1,117
R^2	0.050	0.101	0.107	0.124	0.036	0.045	0.116	0.129
Estimated bias	31.6	30.4	30.9	30.9	31.1	31.2	31.1	30.6
Panel C: Gender Dysphoria (Trans Kids)								
Subject saw list with key statement	0.262***	0.255***	0.260***	0.269***	0.378***	0.376***	0.362***	0.376***
	(0.042)	(0.041)	(0.041)	(0.043)	(0.052)	(0.052)	(0.051)	(0.052)
Observations	1,194	1,194	1,192	1,119	1,194	1,194	1,192	1,119
R^2	0.031	0.089	0.097	0.111	0.043	0.058	0.109	0.141
Estimated bias	14.3	15.0	14.5	13.6	2.7	2.9	4.3	2.9
Controls for:								
Locality fixed effects		✓	✓	✓		✓	✓	✓
Demographic controls		✓	✓	✓		✓	✓	✓
Socio-economic factors			✓	✓			✓	✓
Additional controls and beliefs				✓				✓

Notes: * p < 0.10, ** p < 0.05, *** p < 0.01. OLS estimates. Robust standard errors in parentheses. "Estimated bias" reports the differences (in percentage points) between the estimated percentage of participants who agreed to the key statement in each column and the estimate obtained from the corresponding direct question. *Demographic controls* include subject's age, sex at birth, ethnicity, and sexual orientation. *Socio-economic factors and beliefs* include subject's education level, employment status, income (2 people did not record their income), religiosity, and political affiliation. *Additional controls* include whether subject has a child, whether subject knows someone transgender, whether subject believes that more than 3% of individuals are transgender, urbanicity, marital status, and beliefs about general level of support for puberty blockers for precocious puberty (Panels A-B) or gender dysphoria (Panel C). See also notes in Figure 2.