# The fallacy of objective measurement: The case of gaydar*

Andrew Gelman†

11 Nov 2017

**gaydar**, *n.*

Pronunciation: Brit. `/'geida:/` , U.S. `/'gei,dar/`
Etymology: Blend of gay adj. and radar n.
slang.

An ability, attributed esp. to homosexual people and likened humorously to radar, to identify a (fellow) homosexual person by intuition or by interpreting subtle signals conveyed by appearance or behaviour.
— Oxford English Dictionary (2003)

"Gaydar" colloquially refers to the ability to accurately glean others' sexual orientation from mere observation. But does gaydar really exist? If so, how does it work?

Our research, published recently in the peer-reviewed journal PLoS ONE, shows that gaydar is indeed real and that its accuracy is driven by sensitivity to individual facial features as well as the spatial relationships among facial features.
— Tabak and Zayas (2012b)

## 1. Introduction

Recent news items about "gaydar" reveal the interactions between scientific measurement and definitions. We push against the idea that there is some sort of pure biological measure of perception of sexual orientation. Rather, we argue that the concept of gaydar inherently exists within a social context and that this should be recognized when studying it, and we use this as an example of a more general concern about illusory precision in measurement of social phenomena.

## 2. The mathematics of gaydar

Let $p$ be the proportion of adults in the United States who identify as gay. For simplicity, let us assume that $p = 4\%$ and that everybody is either gay or straight. (It would not be difficult to extend the model to allow for intermediate or other characterizations.) Now suppose that you have a "gaydar" which emits a continuous signal when you study a person, and the reflection of that signal contains a signature that informs you that the person is gay. Or, to put it more prosaically, you process some number of attributes conveyed by a person which enable a probabilistic classification, assuming you have been accurately trained and can correctly adjust for changes in base rate.

To simplify, suppose you are required to compress your continuous gaydar measurement into a binary go/no-go decision. Let $\alpha$ be the probability that a gay person is correctly classified as gay, and let $\beta$ be the probability that a straight person is correctly classified as straight; thus with perfect gaydar, $\alpha = \beta = 1$.

Suppose you classify someone as gay; then the probability that he or she actually is gay is $\Pr(\text{gay} \,|\, \text{classified as gay}) = \frac{p\alpha}{p\alpha+(1-p)(1-\alpha)}$. It is a well known result from conditional probability that when $p$ is low, the misclassification rate is high. For example, if $p = 0.04$ and $\alpha = \beta =$

*We thanks Daniel Simpson for helpful comments.

†Department of Statistics and Department of Political Science, Columbia University, New York

0.9 (which would be an impressive accuracy in many settings), then $\Pr(\text{gay} \,|\, \text{classified as gay}) = \frac{0.04 \cdot 0.9}{0.04 \cdot 0.9 + 0.96 \cdot 0.1} = 0.27$; thus, you would actually be wrong nearly three-quarters of the time. Under this system you would be classifying 13.2% of people as gay—but given that people generally overestimate the proportions of rare groups (including immigrants, ethnic minorities, and gays; see Sides and Citrin, 2007, and Newport, 2015), this is in fact plausible. Srivastava (2011) makes a similar calculation.

## 3. The Tabak and Zayas experiment and later studies

In a study that received extensive media attention, Tabak and Zayas (2012a) assessed the abilities of 24 college-student volunteers at identifying the sexual orientations of 400 self-identified gay or straight people based on photographs of faces which were cleaned up to remove information such as hairstyle.[1] Half the targets in the study were gay and half were straight, and the students correctly identified sexual orientation 60% of the time, which was statistically significantly better than the 50% that would be expected by pure guessing.

Tabak and Zayas write that their research "was the first attempt to determine the roles that featural and configural face processing play in snap judgments of sexual orientation from faces," and it indeed seems to provide a clue about such visual manifestations. But we disagree with their claim that they have shown that "configural face processing significantly contributes to perception of sexual orientation."

To understand our disagreement, consider several aspects of "gaydar" as we understand it, and which are consistent with the dictionary definition above. Gaydar occurs in a social context with information including voice, dress, posture, even topics of conversation; the Tabak and Zayas study removes all cues. Gaydar is relevant in settings where gays are a small fraction of the population (perhaps 3 percent or maybe as high as 10 percent); in contrast, gay people represent 50% of the photos in the experiment under discussion. "Gaydar" has been transmuted from the traditional task of identifying a rare subgroup in a social interaction, to a mechanical binary classification task.

More recently, Wang and Kosinski (2017) performed a similar exercise, this time using a machine learning algorithm to identify faces as gay or straight. This is fine as a classification exercise, and it can be interesting to see what happens to show up in the data (lesbians wear baseball caps!), but their interpretation is way over the top. Its no surprise at all that two groups of people selected from different populations will differ from each other. The ability of an algorithm to classify data from two different samples is taken as an excuse for this sort of thing: " it is unclear whether gay men were less likely to wear a beard because of nature (sparser facial hair) or nurture (fashion). If it is, in fact, fashion (nurture), to what extent is such a norm driven by the tendency of gay men to have sparser facial hair (nature)? Alternatively, could sparser facial hair (nature) stem from potential differences in diet, lifestyle, or environment (nurture)?" As Cohen (2017) and Mattson (2017) note, such speculation is essentially disconnected from the data analysis that is being used as its justification, as well as being blissfully ignorant about the sorts of sampling bias that should caution researchers away from inferring general laws of nature from particular patterns in a particular dataset.

[1] "To minimize the prospect that non-face cues would influence judgments, photographs of men or women with facial alterations or adornments (e.g., scars, eyewear, facial hair, makeup, non-earlobe piercings, etc.) were not included as experimental targets. To maximize consistency across faces, only photographs of White-appearing individuals who self-identified ages of 1829 were included. Using Adobe Photoshop CS3 Extended, research assistants removed hair and ears from each head and converted each image to grayscale (8-bit bitmap format), leaving the final 'face' stimulus."

## 4. Sampling and social context

Both the studies under discussion here measure the perception of sexual orientation in a context-free way. Decontextualization—bringing a phenomenon "into the lab" for careful study—is a characteristic step of scientific measurement, but it can cause problems in fields such as ecology and social science, where context is all.

In particular, we have three concerns with these laboratory studies of gaydar. The easiest concern to state is representativeness. A low frequency of facial hair among gay men who post to a particular dating, or a finding that "sexual orientation is inferred more easily from women's vs. men's faces," may well be telling us more about problems with the samples than about the features of general population. Given that no census or representative sample exists of images of gay people (or, for that matter, straight people), any statistical analysis will always have to deal with the extrapolation problem, and we recommend using some sort of multilevel model that explicitly allows for variation among different subgroups of each population.

Our second concern is the way in which "gaydar," which was originally framed as an aspect of communication *within* the gay community (see the dictionary definition above), has been redefined as a skill applied by the general (thus, mostly straight) population. One can distinguish between "active" gaydar (in which members of a subgroups are sending coded messages to each other) and "passive" gaydar (in which outsiders catch some of these signals even though they are not the intended recipients). *In either case*, the distinctive and noticeable characteristics of the subpopulation are the result of active choices by members of that group, not (as assumed in the two papers under discussion) essential attributes derived from "nature" or "nurture."

By taking gaydar into the lab and putting it under the microscope, these research teams have taken the creative adaptation of an oppressed community, and turned it into an essentialist story of "gender atypicality," moving gay people from the protagonists of the story to the subjects of observation. Again, a certain amount of simplification and objectification is necessary in social science research—but we should be aware of what is lost in these steps.

## 5. The fallacy of objective measurement

The reporting and interpretation of the gaydar experiments suffered from three problems that we feel are important enough to deserve a general discussion.

First, the researchers took a rich real-world phenomenon and abstracted it so much that they removed all its interesting content. "Gaydar" has traditionally existed within a particular social context—a world in which gays are an invisible minority, hiding in plain sight and seeking to be inconspicuous to the general population while communicating with others of their subgroup using various coordination points (see, e.g., Minnelli et al., 1989). How can it make sense to boil this down to the shapes of faces?

Second, the laboratory experiment has a much different base rate than in the real-world setting. It is well known that judgments of uncertainty are contingent on base rate, and this is particularly relevant for a concept such as "gaydar" which arises in a setting in which the challenge is identifying a small minority in a large population.

Third, the report of the laboratory study reduced an (implied) continuous scale to a binary choice. Some people appear clearly gay, others emit some gay signals, while others appear completely straight. "Gaydar" is on a sliding scale and depends on context: again, the traditional goal is to identify gay people who might be signaling their sexual identities to the in-group while staying hidden from the general population, which is quite a bit different from a study such as Wang and

Kosinski (2017), using participants on a dating site who both self-identify as gay and want other people to know this.

We can identify all these problems with what might be called *the fallacy of objective measurement*, the idea that science proceeds by crisp distinctions, with a model being an unambiguous medical diagnosis (a test for strep throat, for example) or a sharp measurement such as the color change in litmus paper. Stripping a phenomenon of its social context, normalizing a base rate to 50%, and seeking an on-off decision: all of these can give the feel of scientific objectivity.

## 6. Discussion

In recent years psychologists have studied correlates of sexual orientation in a variety of ways; see, for example, France, 2007, and Saletan, 2011, for wide-ranging journalistic reviews. The place of homosexuality in our culture has changed dramatically in recent decades, to the extent that concepts such as "gaydar" are changing their meaning.

As noted above, we are not claiming that the Tabak and Zayas (2012) and Wang and Kosinski (2017) experiments are useless. If various aspects of the shapes of faces are (weakly) correlated with sexual orientation, and if untrained volunteers or computer programs can classify such patterns with above-chance accuracy, this is possibly interesting. Some insight might be gained by performing a set of studies comparing other groups, each time using people or computer programs to classify people chosen from two different samples, for example college graduates and non-college graduates, or English people and French people, or drivers license photos in state X and drivers license photos in state Y, or students from college A and students from college B, or baseball players and football players, or people on straight dating site U and people on straight dating site V, or whatever. More generally, reductionism is a characteristic and useful tool of science, understanding a complex phenomenon by breaking it down into simple parts. In this case, however, we think too many steps have been taken in the journey from reality to lab to make any general conclusions about differences between gay and straight people.

The point of this article is not to pick on a small area of psychology research that happened to catch the fancy of the press, or even to criticize larger trends of sensationalism in science and the news media. Rather, we seek to draw attention to the general problem, all to frequent in this era of genetics and functional MRI studies, in which the ideals of scientific precision end up stripping all content from a social phenomenon, leading to nonsensical claims based on predictive accuracy or statistical significance. A social interaction cannot always be measured in a test tube or even a psych lab.

## References


Cohen, P. (2017). On artificially intelligent gaydar. Family Inequality blog, 11 Sep. `https://familyinequality.wordpress.com/2017/09/11/on-artificially-intelligent-gaydar/`

France, D. (2007). The science of gaydar. *New York*, 17 Jun. `http://nymag.com/news/features/33520/`

Hemenway, D. (1997). The myth of millions of annual self-defense gun uses: A case study of survey overestimates of rare events. *Chance* **10** (3), 6–10.

Mattson, G. (2017). Artificial intelligence discovers gayface. Sigh. Gregor Mattson blog, 9 Sep. `https://greggormattson.com/2017/09/09/artificial-intelligence-discovers-gayface/`

Minnelli, L., Tennant, N., Lowe, C., and Mendelsohn, J. (1989). *Results*. New York: Epic Records.

Newport, F. (2015). Americans greatly overestimate percent gay, lesbian in U.S. *Gallup News*, 21 May. http://news.gallup.com/poll/183383/americans-greatly-overestimate-percent-gay-lesbian.aspx

Oxford English Dictionary, third edition (2003). Definition of "gaydar." Oxford University Press. http://www.oed.com/viewdictionaryentry/Entry/247143

Saletan, W. (2011). Read my lisp. *Slate*, 15 Jul. http://www.slate.com/articles/health_and_science/human_nature/2011/07/read_my_lisp.html

Sides, J., and Citrin, J. (2007). European opinion about immigration: The role of identities, interests and information. *British Journal of Political Science* **37**, 477–504.

Srivastava, S. (2011). The precisely fuzzy science of gaydar. The Hardest Science blog, 21 Jul. http://hardsci.wordpress.com/2011/07/21/the-precisely-fuzzy-science-of-gaydar/

Tabak, J., and Zayas, V. (2012a). The roles of featural and configural face processing in snap judgments of sexual orientation. *PLoS One* **7** (5): e36671.

Tabak, J., and Zayas, V. (2012b). The science of 'gaydar.' *New York Times*, 3 Jun, p. SR10.

Wang, Y., and Kosinski, M. (2017). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology.*