Robustness Disparities in Commercial Face Detection

Samuel Dooley
University of Maryland
sdooley1@cs.umd.edu

Tom Goldstein
University of Maryland
tomg@cs.umd.edu

John P. Dickerson
University of Maryland
john@cs.umd.edu

Abstract

Facial detection and analysis systems have been deployed by large companies and critiqued by scholars and activists for the past decade. Critiques that focus on system performance analyze disparity of the system's output, i.e., how frequently is a face detected for different Fitzpatrick skin types or perceived genders. However, we focus on the robustness of these system outputs under noisy natural perturbations. We present the first of its kind detailed benchmark of the robustness of two such systems: Amazon Rekognition and Microsoft Azure. We use both standard and recently released academic facial datasets to quantitatively analyze trends in robustness for each. Qualitatively across all the datasets and systems, we find that photos of individuals who are *older*, *masculine presenting*, of *darker skin type*, or have *dim lighting* are more susceptible to errors than their counterparts in other identities.

1 Introduction

Face detection systems identify the presence and location of faces in images and video. Automated face detection is a core component of myriad systems—including *face recognition technologies* (FRT), wherein a detected face is matched against a database of faces, typically for identification or verification purposes. FRT-based systems are widely deployed [Hartzog, 2020, Derringer, 2019, Weise and Singer, 2020]. Automated face recognition enables capabilities ranging from the relatively morally neutral (e.g., searching for photos on a personal phone [Google, 2021]) to morally laden (e.g., widespread citizen surveillance [Hartzog, 2020], or target identification in warzones [Marson and Forrest, 2021]). Legal and social norms regarding the usage of FRT are evolving [e.g., Grother et al., 2019]. For example, in June 2021, the first county-wide ban on its use for policing [see, e.g., Garvie, 2016] went into effect in the US [Gutman, 2021]. Some use cases for FRT will be deemed socially repugnant and thus be either legally or *de facto* banned from use; yet, it is likely that pervasive use of facial analysis will remain—albeit with more guardrails than are found today [Singer, 2018].

One such guardrail that has spurred positive, though insufficient, improvements and widespread attention is the use of benchmarks. For example, in late 2019, the US National Institute of Standards and Technology (NIST) adapted its venerable Face Recognition Vendor Test (FRVT) to explicitly include concerns for demographic effects [Grother et al., 2019], ensuring such concerns propagate into industry systems. Yet, differential treatment by FRT of groups has been known for at least a decade [e.g., Klare et al., 2012, El Khiyari and Wechsler, 2016], and more recent work spearheaded by Buolamwini and Gebru [2018] uncovers unequal performance at the phenotypic subgroup level. That latter work brought widespread public, and thus burgeoning regulatory, attention to bias in FRT [e.g., Lohr, 2018, Kantayya, 2020].

One yet unexplored benchmark examines the bias present in a system's robustness (e.g., to noise, or to different lighting conditions), both in aggregate and with respect to different dimensions of the population on which it will be used. Many detection and recognition systems are not built in house, instead making use of commercial cloud-based "ML as a Service" (MLaaS) platforms offered by tech giants such as Amazon amd Microsoft. The implementation details of those systems are not

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

exposed to the end user—and even if they were, quantifying their failure modes would be difficult. With this in mind, our **main contribution** is a wide *robustness benchmark* of two commercial-grade face detection systems (accessed via Amazon's Rekognition and Microsoft's Azure face detection APIs). For fifteen types of realistic noise, and five levels of severity per type of noise [Hendrycks and Dietterich, 2019], we test both APIs against images in each of four well-known datasets. Across these more than $5{,}000{,}000$ noisy images, we analyze the impact of noise on face detection performance. Perhaps unsurprisingly, we find that noise decreases overall performance, and that different types of noise impact, in an "unfair" way, cross sections of the population of images (e.g., based on Fitzgerald skin type, age, self-identified gender, and intersections of those dimensions). Our method is extensible and can be used to quantify the robustness of other detection and FRT systems, and adds to the burgeoning literature supporting the necessity of explicitly considering fairness in ML systems with morally-laden downstream uses.

2 Related Work

We briefly overview additional related work in the two core areas addressed by our benchmark: robustness to noise and demographic disparity in facial detection and recognition. That latter point overlaps heavily with the fairness in machine learning literature; for additional coverage of that broader ecosystem and discussion around fairness in machine learning writ large, we direct the reader to survey works due to Chouldechova and Roth [2018] and Barocas et al. [2019].

Demographic effects in facial detection and recognition. The existence of differential performance of facial detection and recognition on groups and subgroups of populations has been explored in a variety of settings. Earlier work [e.g., Klare et al., 2012, O'Toole et al., 2012] focuses on single-demographic effects (specifically, race and gender) in pre-deep-learning face detection and recognition. Buolamwini and Gebru [2018] uncovers unequal performance at the phenotypic subgroup level in, specifically, a gender classification task powered by commercial systems. That work, typically referred to as "Gender Shades," has been and continues to be hugely impactful both within academia and at the industry level. Indeed, Raji and Buolamwini [2019] provide a follow-on analysis, exploring the impact of the Buolamwini and Gebru [2018] paper publicly disclosing performance results, for specific systems, with respect to demographic effects; they find that their named companies (IBM, Microsoft, and Megvii) updated their APIs within a year to address some concerns that were surfaced. Subsequently, the late 2019 update to the NIST FRVT provides evidence that commercial platforms are continuing to focus on performance at the group and subgroup level [Grother et al., 2019]. Further recent work explores these demographic questions with a focus on Indian election candidates [Jain and Parsheera, 2021]. We see our benchmark as adding to this literature by, for the first time, addressing both noise and demographic effects on commercial platforms' face detection offerings.

In this work, we focus on *measuring* the impact of noise on a classification task, like that of Wilber et al. [2016]; indeed, a core focus of our benchmark is to *quantify* relative drops in performance conditioned on an input datapoint's membership in a particular group. We view our work as a *benchmark*, that is, it focuses on quantifying and measuring, decidedly not providing a new method to "fix" or otherwise mitigate issues of demographic inequity in a system. Toward that latter point, existing work on "fixing" unfair systems can be split into three (or, arguably, four [Savani et al., 2020]) focus areas: pre-, in-, and post-processing. Pre-processing work largely focuses on dataset curation and preprocessing [e.g., Feldman et al., 2015, Ryu et al., 2018, Quadrianto et al., 2019, Wang and Deng, 2020]. In-processing often constrains the ML training method or optimization algorithm itself [e.g., Zafar et al., 2017b,a, 2019, Donini et al., 2018, Goel et al., 2018, Padala and Gujar, 2020, Agarwal et al., 2018, Wang and Deng, 2020, Martinez et al., 2020, Diana et al., 2020, Lahoti et al., 2020], or focuses explicitly on so-called fair representation learning [e.g., Adeli et al., 2021, Dwork et al., 2012, Zemel et al., 2013, Edwards and Storkey, 2016, Madras et al., 2018, Beutel et al., 2017, Wang et al., 2019]. Post-processing techniques adjust decisioning at inference time to align with quantitative fairness definitions [e.g., Hardt et al., 2016, Wang et al., 2020].

Robustness to noise. Quantifying, and improving, the robustness to noise of face detection and recognition systems is a decades-old research challenge. Indeed, mature challenges like NIST's Facial Recognition Vendor Test (FRVT) have tested for robustness since the early 2000s [Phillips et al., 2007]. We direct the reader to a comprehensive introduction to an earlier robustness challenge due to NIST [Phillips et al., 2011]; that work describes many of the specific challenges faced by

Figure 1: Our benchmark consists of 5,066,312 images of the 15 types of algorithmically generated corruptions produced by ImageNet-C. We use data from four datasets (Adience, CCD, MIAP, and UTKFace) and present examples of corruptions from each dataset here.

face detection and recognition systems, often grouped into Pose, Illumination, and Expression (PIE). It is known that commercial systems still suffer from degradation due to noise [e.g., Hosseini et al., 2017]; none of this work also addresses the intersection of noise with fairness, as we do. Recently, *adversarial* attacks have been proposed that successfully break commercial face recognition systems [Shan et al., 2020, Cherepanova et al., 2021]; we note that our focus is on *natural* noise, as motivated by Hendrycks and Dietterich [2019] by their ImageNet-C benchmark. Literature at the intersection of adversarial robustness and fairness is nascent and does not address commercial platforms [e.g., Singh et al., 2020, Nanda et al., 2021]. To our knowledge, our work is the first systematic benchmark for commercial face detection systems that addresses, comprehensively, noise and its differential impact on (sub)groups of the population.

3 Experimental Description

Datasets and Protocol. This benchmark uses four datasets to evaluate the robustness of Amazon AWS and Microsoft Azure's face detection systems. They are described below.

The Open Images Dataset V6 – Extended; More Inclusive Annotations for People (**MIAP**) dataset [Schumann et al., 2021] was released by Google in May 2021 as a extension of the popular, permissive-licensed Open Images Dataset specifically designed to improve annotations of humans. For each image, every human is exhaustively annotated with bounding boxes for the entirety of their person visible in the image. Each annotation also has perceived gender (Feminine/Masculine/Unknown) presentation and perceived age (Young, Middle, Old, Unknown) presentation.

The Casual Conversations Dataset (**CCD**) [Hazirbas et al., 2021] was released by Facebook in April 2021 under limited license and includes videos of actors. Each actor consented to participate in an ML dataset and provided their self-identification of age and gender (coded as Female, Male, and Other), each actor's skin type was rated on the Fitzpatrick scale [Fitzpatrick, 1988], and each video was rated for its ambient light quality. For our benchmark, we extracted one frame from each video.

The **Adience** dataset [Eidinger et al., 2014] under a CC license, includes cropped images of faces from images "in the wild". Each cropped image contains only one primary, centered face, and each face is annotated by an external evaluator for age and gender (Female/Male). The ages are reported as member of 8 age range buckets: 0-2; 3-7; 8-14; 15-24; 25-35; 36-45; 46-59; 60+.

Finally, the **UTKFace** dataset [Zhang et al., 2017] under a non-commercial license, contains images with one primary subject and were annotated for age (continuous), gender (Female/Male), and ethnicity (White/Black/Asian/Indian/Others) by an algorithm, then checked by human annotators.

For each of the datasets, we randomly selected a subset of images for our evaluation in order to cap the number of images from each intersectional identity at $1{,}500$ as an attempt to reduce the effect of highly imbalanced datasets. We include a total of $66{,}662$ images with $14{,}919$ images from Adience; $21{,}444$ images from CCD; $8{,}194$ images from MIAP; and $22{,}105$ images form UTKFace. The full breakdown of totals of images from each group can be found in Section A.1.

Each image was corrupted a total of 75 times, per the ImageNet-C protocol with the main 15 corruptions each with 5 severity levels. Examples of these corruptions can be seen in Figure 1. This resulted in a total of $5{,}066{,}312$ images (including the original clean ones) which were each passed through the AWS and Azure face analysis systems. A detailed description of which API settings were selected can be found in Appendix C. The API calls were conducted between 19 May and 29 May 2021. Images were processed and stored within AWS's cloud using S3 and EC2. The total cost of the experiments was $\$9{,}887.17$ and a breakdown of costs can be found in Appendix D.

Evaluation Metrics. Given that we aim is to study how corruptions to an image alter the commercial interpretation of that image, we valuate the error of the face systems. Additionally, none of the chosen datasets have ground truth face bounding boxes. Therefore, we can use the response from the clean image as a ground truth of sorts. Specifically, we take as ground truth the number of faces in an clean image and compare that to the number of faces detected in a corrupted image.

Our main metric is the relative error in the number of faces a system detects after corruption; this metric has been used in other facial processing benchmarks [Jain and Parsheera, 2021]. Measuring error in this way is in some sense incongruous with the object detection nature of the APIs. However, none of the data in our datasets have bounding boxes for each face. This means that we cannot calculate precision metrics as one would usually do with other detection tasks. To overcome this, we hand annotated bounding boxes for each face in 772 random images from the dataset. We then calculated per-image precision scores (with an IoU of 0.5) and per-image relative error in face counts and we find a Pearson's correlation of 0.91 (with $p < 0.001$). This high correlation indicates that the proxy is sufficient to be used in this benchmark in the absence of fully annotated bounding boxes.

This error is calculated for each image. The way in which this works is that we first pass every clean, uncorrupted image through the commercial system's API. Then, we measure the number of detected faces, i.e., length of the system's response, and treat this number as the ground truth. Subsequently, we compare the number of detected faces for a corrupted version of that image. If the two face counts are not the same, then we call that an error. We refer to this as the *relative corruption error*. For each clean image, i, from dataset d, and each corruption c which produces a corrupted image $\hat{i}_{c,s}$ with severity s, we compute the relative corruption error for system r as

$$rCE^{d,r}_{c,s}(\hat{i}_{c,s}) := \begin{cases} 1, & \text{if } l_r(i) \neq l_r(\hat{i}_{c,s}) \\ 0, & \text{if } l_r(i) = l_r(\hat{i}_{c,s}) \end{cases}$$

where l_r computes the number of detected faces, i.e., length of the response, from face detection system r when given an image. Often the super- and subscripts are omitted when they are obvious from context.

Our main metric, relative error, aligns with that of the ImageNet-C benchmark. We report mean relative corruption error ($mrCE$) defined as taking the average of rCE across some relative set of categories. In our experiments, depending on the context, we might have any of the following categories: face systems, datasets, corruptions, severities, age presentation, gender presentation, Fitzpatrick rating, and ambient lighting. For example, we might report the relative mean corruption error when averaging across demographic groups; the mean corruption error for Azure on the UTK dataset for each age group a is $mrCE_a = \frac{1}{15}\frac{1}{5}\sum_{c,s} rCE^{UTK,Azure}_{c,s,a}$. The subscripts on $mrCE$ will be omitted when it is obvious what their value is in whatever context they are presented.

Finally, we will also investigate the significance of whether the $mrCE$ for two groups are equal. For example, our first question is whether the two commercial systems (AWS and Azure) have comparable $mrCE$ overall. To do this, we will report the raw $mrCE$; these frequency or empiric probability statistics offer much insight into the likelihood of error. But we also indicate the statistical significance at $\alpha = 0.05$ determined by logistic regressions for the appropriate variables and interactions. For each claim of significance, regression tables can be found in the appendix. Accordingly, we discuss the odds or odds ratio of relevant features. See Appendix B for a detailed example. Finally, each

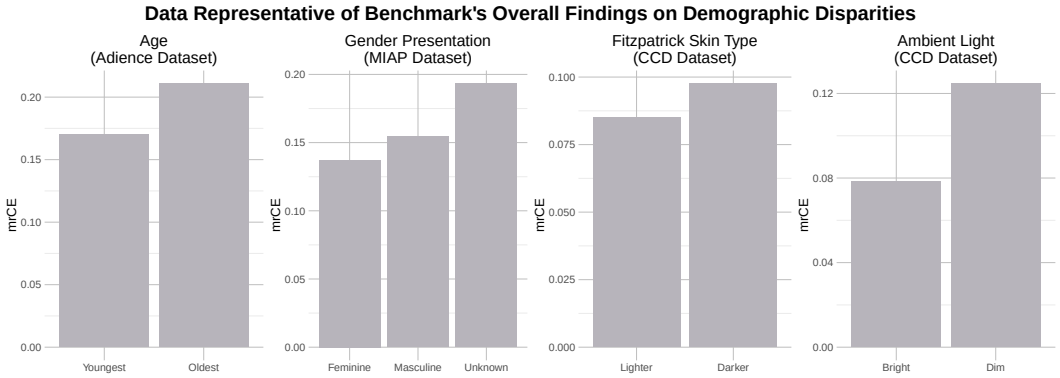

Figure 2: There are disparities in all of the demographics included in this study; we show representative evidence for each demographic on different datasets. On the left, we see (using Adience as an exemplar) that the oldest two age groups are roughly 25% more error prone than the youngest two groups. Using MIAP as an exemplar, masculine presenting subjects are 20% more error prone than feminine. On the CCD dataset, we find that individuals with Fitzpatrick scales IV-VI have a roughly 25% higher chance of error than lighter skinned individuals. Finally, dimly lit individuals are 60% more likely to have errors.

claim we make for an individual dataset or service is backed up with statistical rigor through the logistic regressions. Each claim we make across datasets is done by looking at the trends in each dataset and are inherently qualitative.

What is not included in this study. There are three main things that this benchmark does not address. First, we do not examine cause and effect. We report inferential statistics without discussion of what generates them. Second, we only examine the types of algorithmicaly generated natural noise present in the 15 corruptions. We speak narrowly about robustness to these corruptions or perturbations. We explicitly do not study or measure robustness to other types of changes to images, for instance adversarial noise, camera dimensions, etc. Finally, we do not investigate algorithmic training. We do not assume any knowledge of how the commercial system was developed or what training procedure or data were used.

Social Context. The central analysis of this benchmark relies on socially constructed concepts of gender presentation and the related concepts of race and age. While this benchmark analyzes phenotypal versions of these from metadata on ML datasets, it would be wrong to interpret our findings absent a social lens of what these demographic groups mean inside a society. We guide the reader to Benthall and Haynes [2019] and Hanna et al. [2020] for a look at these concepts for race in machine learning, and Hamidi et al. [2018] and Keyes [2018] for similar looks at gender.

4 Benchmark Results

We now report the main results of our benchmark, a synopsis of which is in Figure 2. Overall, we find that photos of individuals who are *older*, *masculine presenting*, *darker skinned*, or are *dimly lit* are more susceptible to errors than their counterparts. We come to these qualitative conclusions by quantitatively examining the trends of each dataset for each demographic. All four datsaets have age and gender labels. We see the bias against older individuals across all datasets. The bias against masculine presenting individuals is present in all datasets except UTKFace (which shows no bias). Skin type and lighting labels are only present in one dataset, CCD.

Below is a more detailed analysis with additional supporting tables and figures in the Appendix.

4.1 System Performance

Overall, AWS has fewer errors than Azure on corrupted data though the magnitude of the difference is small. The $mrCE$ for AWS is 12.298% whereas Azure is 12.338%, or 3% higher, but this is a Simpson's Paradox because when we look at each dataset, we see further nuance.

We plot the CTRs for each dataset and service in Figure 3; the difference between services is statistically significant for each dataset. For the Adience and MIAP datasets, Azure performs better

Figure 3: Observe that AWS is more robust on CCD and UTK and Azure is more robust on Adience and MIAP.

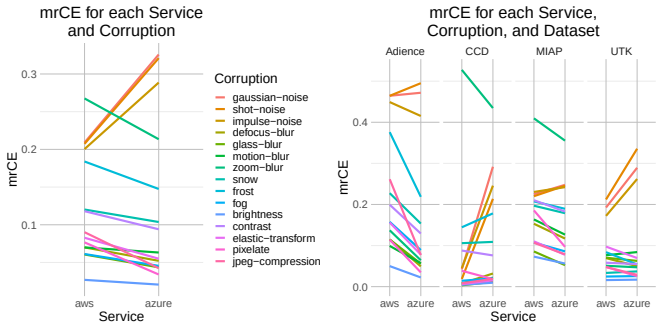

Figure 4: A comparison of $mrCE$ for each commercial system and dataset where each line represents one of the 15 types of corruptions. (Left) depicts the robustness across all datasets whereas (right) depicts this for each dataset separately.

than AWS. On Adience, Azure's $mrCE$ is 16.1% whereas AWS has $mrCE$ of 22.6%. The magnitude is less on MIAP; Azure has 15.6% and AWS has 17.9%.

Conversely, on the CCD and UTK dataset, Azure outperforms AWS. For the CCD dataset, Azure performs 60% worse than AWS (AWS $mrCE$ of 7.1% compared to Azure's 11.3%). The magnitude is less on UTKFace; AWS has 8.4% whereas Azure has 9.6%.

4.2 Noise corruptions are the most difficult

Recall that the ImageNet-C corruptions are broken into four different types: noise, blur, weather, and digital corruptions. We observe that the noise corruptions prove to be some of the most difficult corruptions for the commercial systems to handle. From Figure 4, we observe that in the AWS system, the three noise corruptions have the the second, third, and fourth most difficult corruptions (behind zoom blur). However, they are markedly the most difficult corruptions for Azure to handle. On the otherhand, Azure outperforms AWS on every other corruption. The difficulty of the noise corruptions echos that documented in the ImageNet-C experiments, though the comparative magnitude of the difficulty for these systems is significantly higher than what is previously documented.

When we examine the differences in the performance for each corruption across the different datasets, we see a continuation of the theme that the noise corruptions have relatively high $mrCE$. In every instance except one, Azure performs worse on the noise corruptions than AWS. For both commercial systems on Adience, the $mrCE$ values for the noise corruptions are above 40%. However, Azure preforms better than AWS on all other corruptions on the Adience Dataset.

The zoom blur corruption proves particularly difficult on the CCD and MIAP datasets, though Azure is significantly better than AWS (CCD: 52.7% for AWS and 43.5% for Azure; MIAP: 41.0% for AWS and 35.5% for Azure). We also note that all corruptions for all datasets and commercial systems are significantly differently from zero.

4.2.1 Comparison to ImageNet-C results

Even though Hendrycks and Dietterich [2019] worked with the ImageNet dataset, we compare the findings from their paper to our experiments. We recreate Figure 3 from their paper with more current results for recent models since their paper was published, as well as the addition of our findings for AWS and Azure's face detection on our data; see Figure 8. This figure reproduces their metric, mean corruption error and relative mean corruption error. These differ from our metrics as they are defined as the raw error for each corruption, but normalized against the performance of AlexNet from the original paper. This is done so as to compare different models more fairly. The figure also shows the relative mean corruption error which is the difference between the raw error for each corruption and the raw error for the clean data. From this figure, we can conclude that our results are very highly in-line with the predictions from the previous data. This indicates that, even with highly accurate models, accuracy is a strong predictor of robustness.

We also examined the corruption-specific differences between our findings (with face data) and that of the original paper (with ImageNet data). We find that while facial datasets are most susceptible to

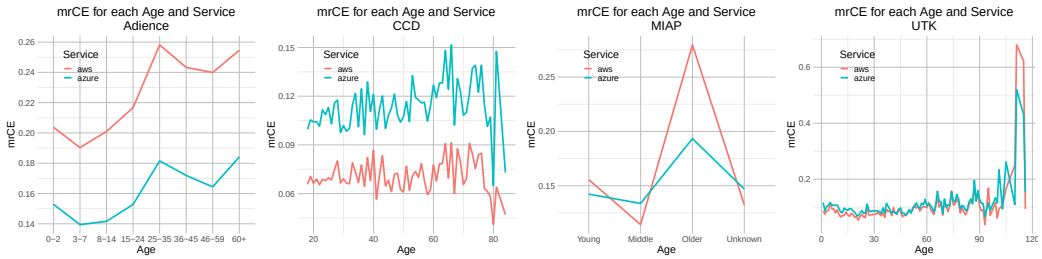

Figure 5: Each figure depicts the $mrCE$ across ages. Each line depicts a commercial system (AWS is above Azure for Adience and MIAP). Age is a categorical variable for Adience and MIAP but a numeric for CCD and UTKFace. Observe the general trend of increased errors for older individuals.

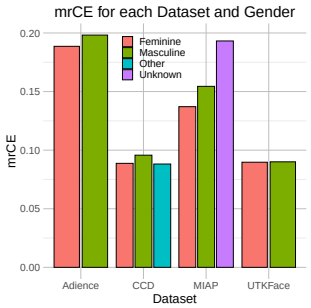

Figure 6: Observe that on all datasets, except for UTKFace, feminine presenting individuals are more robust than masculine.

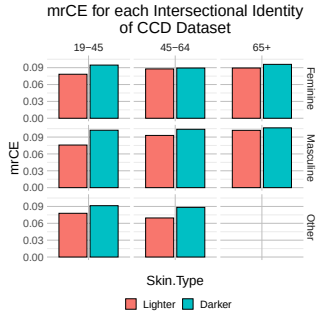

Figure 7: In all intersectional identities, except for 45-64 females, darker skinned individuals are less robust than those who are lighter skinned.

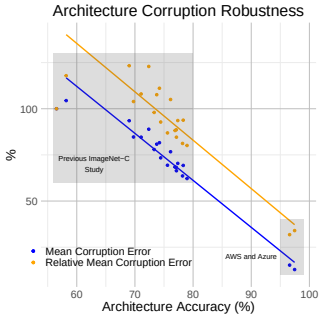

Figure 8: Recreation of Figure 3 from Hendrycks and Dietterich [2019] with new results since their paper and the addition of our findings.

noise corruptions, zoom blur, weather, etc, the ImageNet datasets are generally uniformly susceptible to corruptions with blurs and digital corruptions being the most difficult for them. This indicates that the face data have qualitative differences in their robustness susceptibility, indicating a need for further study.

4.3 Errors increase on older subjects

We observe a significant impact of age on $mrCE$. See Figure 5. In every dataset and every commercial system, we see that older subjects have significantly higher error rates. Recall that all four datasets have age metadata. Adience and MIAP have such data in groups. CCD and UTKFace have age data as a continuous variable.

On the Adience dataset, there is an interesting behavior where the second and third youngest age groups have the best performance with increases for younger and older age groups. There is then a spike in errors in the 25-35 age group which falls off slightly for the 36-59 groups and finally increases again for the oldest 60+ group. These two maximal groups have nearly 1:4 odds of error. This is compared to the youngest group which has 30% better odds (3:15).

For the MIAP dataset, the age disparity is very pronounced. Like the Adience dataset, we see a decrease in the likelihood of error moving from the youngest to the middle ages. However, we see a very large increase for the Oldest individuals. In AWS for instance, we see a 145% increase in error.

The CCD and UTKFace datasets have numeric age. Analyzing the regressions indicates that for every increase of 10 years, there is a 2.3% increase in the likelihood of error on the CCD data and 2.7% increase for UTKFace data. In Appendix E.4, we explore the interaction of Age and the corruptions.

4.4 Masculine presenting individuals have more errors than feminine presenting

Across all datasets except UTKFace, we find that feminine presenting individuals have lower errors than masculine presenting individuals. See Figure 6. On Adience, feminine individuals have 18.8% $mrCE$ whereas masculine have 19.8%. On CCD, the $mrCE$s are 8.9% and 9.6% respectively. On

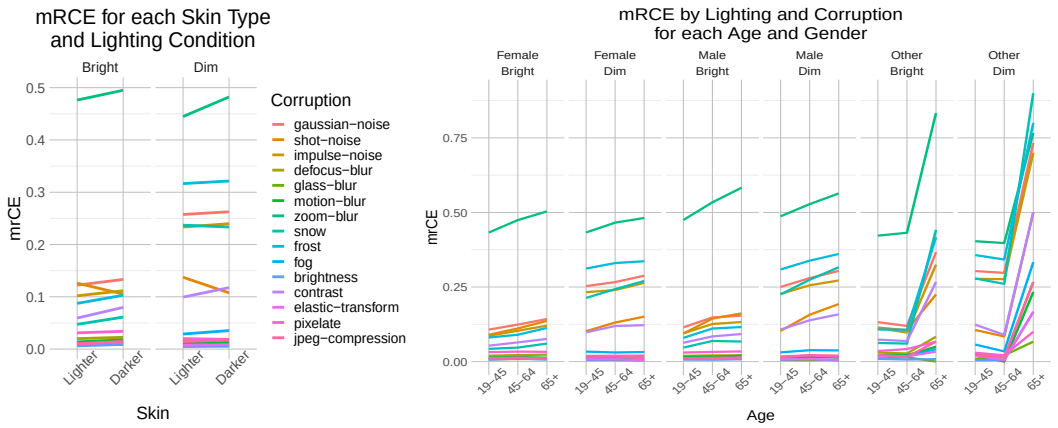

Figure 9: (Left) $mrCE$ is plotted for each corruption by the intersection of lighting condition and skin type. (Right) the same is plotted by the intersection of age, gender, and lighting. Observe that for both skin types, all genders, and all ages, the dimly lit environment increases the error rates. Motion blur is the least robust corruption with frost, the three noises, and snow being the next worst across most intersectional identities.

the MIAP dataset, the $mrCE$ values are 13.7% and 15.4% respectively. On the UTKFace, both gender presentations have around 9.0% $mrCE$ (non statistically significant difference).

Stepping outside the gender binary, we have two insights into this from these data. In the CCD dataset, the subjects were asked to self-identify their gender. Two individuals selected Other and 62 others did not provide a response. Those two who chose outside the gender binary have a $mrCE$ of 4.9%. When we include those individuals without gender labels, their $mrCE$ is 8.8% and not significantly different from the feminine presenting individuals.

The other insight comes from the MIAP dataset where subjects were rated on their perceived gender presentation by crowdworkers; options were “Predominantly Feminine", “Predominantly Masculine", and "Unknown". For those “Unknown", the overall $mrCE$ is 19.3%. The creators of the dataset automatically set the gender presenation of those with an age presentation of “Young" to be “Unknown". The $mrCE$ of those annotations which aren’t “Young" and have an “Unknown" gender presentation raises to 19.9%. One factor that might contribute to this phenomenon is that individuals with an “Unknown” gender presentation might have faces that are occluded or are small in the image. Further work should be done to explore the causes of his discrepancy. In Appendix E.3, we explore the interaction of Gender and the corruptions.

4.5 Dark skinned subjects have more errors across age and gender identities

We analyze data from the CCD dataset which has ratings for each subject on the Fitzpatrick scale. As is customary in analyzing these ratings, we split the six Fitzpatrick values into two: Lighter (for ratings I-III) and Darker for ratings (IV-VI). The main intersectional results are reported in Figure 7.

The overall $mrCE$ for lighter and darker skin types are 8.5% and 9.7% respectively, a 15% increase for the darker skin type. We also see a similar trend in the intersectional identities available in the CCD metadata (age, gender, and skin type). We see that in every identity (except for 45-64 year old and Feminine) the darker skin type has statistically significant higher error rates. This difference is particularly stark in 19-45 year old, masculine subjects. We see a 35% increase in errors for the darker skin type subjects in this identity compared to those with lighter skin types. For every 20 errors on a light skinned, masculine presenting individual between 18 and 45, there are 27 errors for dark skinned individuals of the same category.

4.6 Dim lighting conditions has the most severe impact on errors

Using lighting condition information from the CCD dataset, we observe the $mrCE$ is substantially higher in dimly lit environments: 12.5% compared to 7.8% in bright environments. See Figure 9.

Across the board, we generally see that the disparity in demographic groups decreases between bright and dimly lit environments. For example, the odds ratio between dark and light skinned subjects is 1.09 for bright environments, but decreases to 1.03 for dim environments. This is true for age groups

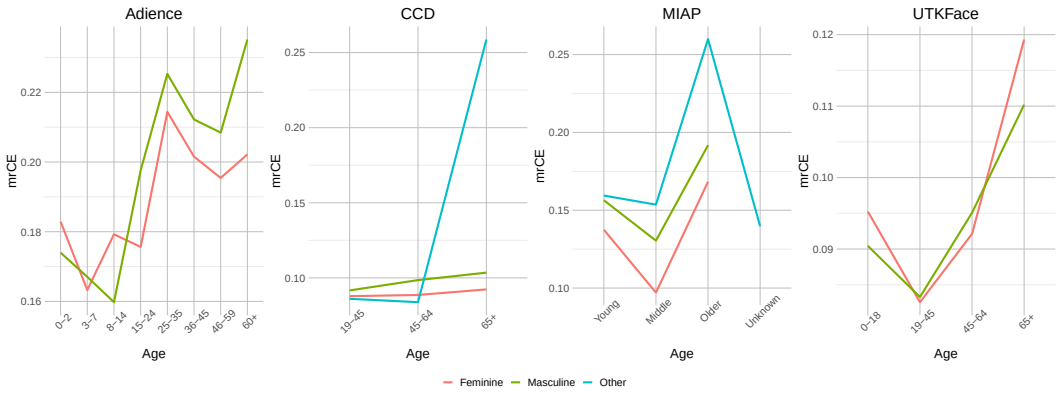

Figure 10: For each dataset, the $mrCE$ is plotted across age groups. Each gender is represented and indicates how gender disparities change across the age groups.

(e.g., odds ratios 1.183 (bright) vs 1.127 (dim) for 45-64 compared to 19-45; 1.138 (bright) vs 1.060 (dim) for Males compared to Females). This is not true for individuals with gender identities as Other or omitted – the disparity increases (1.104 (bright) vs 1.173 (dim) with Females as the reference).

In Figure 9 we observe the lighting differences for different intersectional identities across corruptions. We continue to see zoom blur as the most challenging corruption. Interestingly, the noise and some weather corruptions have a large increase in their errors in dimly lit environments across intersectional identities whereas many of the other corruptions do not.

4.7 Older subjects have higher gender error disparities

We plot in Figure 10 the $mrCE$ for each dataset across age with each gender group plotted separately. From this, we can note that on the CCD and MIAP dataset, the masculine presenting group is always less robust than the feminine. On the CCD dataset, the disparity between the two groups increases as the age increases (odds ratio of 1.048 for 19-45 raises to 1.135 for 65+). On the MIAP dataset, the odds ratio is greatest between masculine and feminine for the middle age group (1.395). The disparities between the ages also increases from feminine to masculine to unknown gender identities.

On the Adience and UTKFace datasets, we see that the feminine presenting individuals sometime have higher error rates than masculine presenting subjects. Notably, the most disparate errors in genders on these datasets occurs at the oldest categories, following the trend from the other datasets.

5 Gender and Age Estimation Analysis

We briefly overview results from evaluating AWS's age and gender estimation commercial systems. The detection model we evaluated for Azure does not provide age and gender estimates. Further analysis can be found in Appendices F and G.

5.1 Gender estimation is at least twice as susceptible to corruptions as face detection

The use of automated gender estimates in ML is a controversial topic. Trans and gender queer individuals are often ignored in ML research, though there is a growing body of research that aims to use these technologies in an assistive way as well [e.g., Ahmed, 2019, Chong et al., 2021]. To evaluate gender estimation, we only use CCD as the subjects of these photos voluntarily identified their gender. We omit from the analysis any individual who either did not choose to give their gender or fall outside the gender binary because AWS only estimates Male and Female.

AWS misgenders 9.1% of the clean images but 21.6% of the corrupted images. Every corruption performs worse on gender estimation than $mrCE$. Two corruptions (elastic transform and glass blur) do not have statistically different errors from the clean images. All the others do, with the most significant being zoom blur, Gaussian noise, impulse noise, snow, frost, shot noise, and contrast. Zoom blur's probability of error is 61% and Gaussian noise is 32%. This compares to $mrCE$ values of 43% and 29% respectively. See Appendix F for further analysis.

5.2 Corrupted images error in their age predictions by 40% more than clean images

To estimate Age, AWS returns an upper and lower age estimation. Following their own guidelines on face detection,[1] we use the mid-point of these numbers as a approximate estimate. On average, the estimation is 8.3 years away from the actual age of the subject for corrupted data, this compares to 5.9 years away for clean data. See Appendix G for further analysis.

6 Conclusion

This benchmark has evaluated two leading commercial facial detection and analysis systems for their robustness against common natural noise corruptions. Using the 15 ImageNet-C corruptions, we measured the relative mean corruption error as measured by comparing the number of faces detected in a clean and corrupted image. We used four academic datasets which included demographic detail. Adience, MIAP, and UTKFace have perceived age and gender metadata. CCD has subject provided age and gender responses as well as external ratings of skin type and ambient lighting conditions.

We observed through our analysis that there are significant demographic disparities in the likelihood of error on corrupted data. We found that older individuals, masculine presenting individuals, those with darker skin types, or in photos with dim ambient light all have higher errors ranging from 20-60%. We also investigated questions of intersectional identities finding that darker males have the highest corruption errors. As for age and gender estimation, corruptions have a significant and sizeable impact on the system's performance; gender estimation is more than twice as bad on corrupted images as it is on clean images; age estimation is 40% worse on corrupted images.

Future work could explore other metrics for evaluating face detection systems when ground truth bounding boxes are not present. While we considered the length of response on clean images to be ground truth, it could be viable to treat the clean image's bounding boxes as ground truth and measure deviations therefrom when considering questions of robustness. Of course, this would require a transition to detection-based metrics like precision, recall, and F-measure.

We do not explore questions of causation in this benchmark. We do not have enough different datasets or commercial systems to probe this question through regressions or mixed effects modeling. We do note that there is work that examines causation questions with such methods like that of [Best-Rowden and Jain, 2017] and [Cook et al., 2019]. With additional data and under similar benchmarking protocols, one could start to examine this question. However, the black-box nature of commercial systems presents unique challenges to this endeavor.

[1] `https://docs.aws.amazon.com/rekognition/latest/dg/guidance-face-attributes.html`

References

E. Adeli, Q. Zhao, A. Pfefferbaum, E. V. Sullivan, L. Fei-Fei, J. C. Niebles, and K. M. Pohl. Representation learning with statistical independence to mitigate bias. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2513–2523, 2021.

A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 60–69, 2018. URL `http://proceedings.mlr.press/v80/agarwal18a.html`.

A. A. Ahmed. Bridging social critique and design: Building a health informatics tool for transgender voice. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–4, 2019.

S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. `http://www.fairmlbook.org`.

S. Benthall and B. D. Haynes. Racial categories in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 289–298, 2019.

L. Best-Rowden and A. K. Jain. Longitudinal study of automatic face recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(1):148–162, 2017.

A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81, pages 77–91, 2018. URL `http://proceedings.mlr.press/v81/buolamwini18a.html`.

V. Cherepanova, M. Goldblum, H. Foley, S. Duan, J. P. Dickerson, G. Taylor, and T. Goldstein. Lowkey: leveraging adversarial attacks to protect social media users from facial recognition. In *International Conference on Learning Representations (ICLR)*, 2021.

T. Chong, N. Maudet, K. Harima, and T. Igarashi. Exploring a makeup support system for transgender passing based on automatic gender recognition. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.

A. Chouldechova and A. Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury. Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 1(1):32–41, 2019.

W. Derringer. A surveillance net blankets china's cities, giving police vast powers. *The New York Times*, Dec. 17 2019. URL `https://www.nytimes.com/2019/12/17/technology/china-surveillance.html`.

E. Diana, W. Gill, M. Kearns, K. Kenthapadi, and A. Roth. Convergent algorithms for (relaxed) minimax fairness. *arXiv preprint arXiv:2011.03108*, 2020.

M. Donini, L. Oneto, S. Ben-David, J. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 2796–2806, 2018.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450311151. doi: 10.1145/2090236.2090255. URL `https://doi.org/10.1145/2090236.2090255`.

H. Edwards and A. J. Storkey. Censoring representations with an adversary. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.05897.

E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on Information Forensics and Security*, 9(12):2170–2179, 2014.

H. El Khiyari and H. Wechsler. Face verification subject to varying (age, ethnicity, and gender) demographics using deep learning. *Journal of Biometrics and Biostatistics*, 7(323):11, 2016.

M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Knowledge Discovery and Data Mining*, pages 259–268, 2015.

T. B. Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124(6):869–871, 1988.

C. Garvie. *The perpetual line-up: Unregulated police face recognition in America*. Georgetown Law, Center on Privacy & Technology, 2016.

N. Goel, M. Yaghini, and B. Faltings. Non-discriminatory machine learning through convex fairness criteria. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018. URL https://ojs.aaai.org/index.php/AAAI/article/view/11662.

Google. How google uses pattern recognition to make sense of images. https://policies.google.com/technologies/pattern-recognition?hl=en-US, 2021. Accessed: 2021-06-07.

P. Grother, M. Ngan, and K. Hanaoka. *Face Recognition Vendor Test (FVRT): Part 3, Demographic Effects*. National Institute of Standards and Technology, 2019.

D. Gutman. King County Council bans use of facial recognition technology by Sheriff's Office, other agencies. *The Seattle Times*, June 2021. URL https://www.seattletimes.com/seattle-news/politics/king-county-council-bans-use-of-facial-recognition-technology-by-sheriffs-office-other-agencies/.

F. Hamidi, M. K. Scheuerman, and S. M. Branham. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–13, 2018.

A. Hanna, E. Denton, A. Smart, and J. Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 501–512, 2020.

M. Hardt, E. Price, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, pages 3315–3323, 2016. URL https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

W. Hartzog. The secretive company that might end privacy as we know it. *The New York Times*, Jan. 18 2020. URL https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.

C. Hazirbas, J. Bitton, B. Dolhansky, J. Pan, A. Gordo, and C. C. Ferrer. Towards measuring fairness in ai: the casual conversations dataset. *arXiv preprint arXiv:2104.02821*, 2021.

D. Hendrycks and T. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. 2019.

H. Hosseini, B. Xiao, and R. Poovendran. Google's cloud vision API is not robust to noise. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pages 101–105. IEEE, 2017.

G. Jain and S. Parsheera. 1.4 billion missing pieces? auditing the accuracy of facial processing tools on indian faces. *First Workshop on Ethical Considerations in Creative applications of Computer Vision*, 2021.

S. Kantayya. Coded bias, 2020. Feature-length documentary.

O. Keyes. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proceedings of the ACM on human-computer interaction*, 2(CSCW):1–22, 2018.

B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6): 1789–1801, 2012.

P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, and E. H. Chi. Fairness without demographics through adversarially reweighted learning. *arXiv preprint arXiv:2006.13114*, 2020.

S. Lohr. Facial recognition is accurate, if you're a white guy. *New York Times*, 9, 2018.

D. Madras, E. Creager, T. Pitassi, and R. S. Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3381–3390. PMLR, 2018. URL http://proceedings.mlr.press/v80/madras18a.html.

J. Marson and B. Forrest. Armed low-cost drones, made by turkey, reshape battlefields and geopolitics. *The Wall Street Journal*, Jun 2021. URL https://www.wsj.com/articles/armed-low-cost-drones-made-by-turkey-reshape-battlefields-and-geopolitics-11622727370.

N. Martinez, M. Bertran, and G. Sapiro. Minimax pareto fairness: A multi objective perspective. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 6755–6764, 2020. URL http://proceedings.mlr.press/v119/martinez20a.html.

V. Nanda, S. Dooley, S. Singla, S. Feizi, and J. P. Dickerson. Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 466–477, 2021.

A. J. O'Toole, P. J. Phillips, X. An, and J. Dunlop. Demographic effects on estimates of automatic face recognition performance. *Image and Vision Computing*, 30(3):169–176, 2012.

M. Padala and S. Gujar. Fnnc: Achieving fairness through neural networks. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2277–2283. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/315. URL https://doi.org/10.24963/ijcai.2020/315.

P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. Frvt 2006 and ice 2006 large-scale results. *National Institute of Standards and Technology, NISTIR*, 7408(1):1, 2007.

P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. S. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer. An introduction to the good, the bad, & the ugly face recognition challenge problem. In *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pages 346–353. IEEE, 2011.

N. Quadrianto, V. Sharmanska, and O. Thomas. Discovering fair representations in the data domain. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8227–8236. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00842. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Quadrianto_Discovering_Fair_Representations_in_the_Data_Domain_CVPR_2019_paper.html.

I. D. Raji and J. Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435, 2019.

H. J. Ryu, H. Adam, and M. Mitchell. Inclusivefacenet: Improving face attribute detection with race and gender diversity. *arXiv preprint arXiv:1712.00193*, 2018.

Y. Savani, C. White, and N. S. Govindarajulu. Intra-processing methods for debiasing neural networks. In *Proceedings of Advances in Neural Information Processing Systems*, 2020.

C. Schumann, C. R. Pantofaru, S. Ricco, U. Prabhu, and V. Ferrari. A step toward more inclusive people annotations for fairness. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2021.

S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, pages 1589–1604, 2020.

N. Singer. Microsoft urges congress to regulate use of facial recognition. *The New York Times*, 2018.

R. Singh, A. Agarwal, M. Singh, S. Nagpal, and M. Vatsa. On the robustness of face recognition algorithms against attacks and bias. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13583–13589, 2020.

M. Wang and W. Deng. Mitigating bias in face recognition using skewness-aware reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9322–9331, 2020.

T. Wang, J. Zhao, M. Yatskar, K.-W. Chang, and V. Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5310–5319, 2019.

Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation, 2020.

K. Weise and N. Singer. Amazon pauses police use of its facial recognition software. *The New York Times*, Jul. 10 2020. URL `https://www.nytimes.com/2020/06/10/technology/amazon-facial-recognition-backlash.html`.

M. J. Wilber, V. Shmatikov, and S. Belongie. Can we still avoid automatic face detection? In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016.

M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact. *Proceedings of the 26th International Conference on World Wide Web*, Apr 2017a. doi: 10.1145/3038912.3052660. URL `http://dx.doi.org/10.1145/3038912.3052660`.

M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pages 962–970. PMLR, 2017b. URL `http://proceedings.mlr.press/v54/zafar17a.html`.

M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019. URL `http://jmlr.org/papers/v20/18-262.html`.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL `http://proceedings.mlr.press/v28/zemel13.html`.

Z. Zhang, Y. Song, and H. Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 3.
   (b) Did you mention the license of the assets? [Yes] See Section 3.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See Section 3, and references within each of the papers that introduce the datasets that we use and the noise models that we use.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Section 3.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Evaluation Information

A.1 Image Counts

For each dataset, we selected no more than 1,500 images from any intersectional group. The final tallies of how many images from each group can be found in Tables 1, 2, 3, and 4.

A.2 Corruption information

We evaluate 15 corruptions from Hendrycks and Dietterich [2019]: Gaussian noise, shot noise, impulse noise, defocus blur, glass blur, motion blur, zoom blur, snow, frost, fog, brightness, contrast, elastic transforms, pixelation, and jpeg compressions. Each corruption is described in the Hendrycks and Dietterich [2019] paper as follows:

The first corruption type is Gaussian noise. This corruption can appear in low-lighting conditions. Shot noise, also called Poisson noise, is electronic noise caused by the discrete nature of light itself. Impulse noise is a color analogue of salt-and-pepper noise and can be caused by bit errors. Defocus blur occurs when an image is out of focus. Frosted Glass Blur appears with "frosted glass" windows or panels. Motion blur appears when a camera is moving quickly. Zoom blur occurs when a camera moves toward an object rapidly. Snow is a visually obstructive form of precipitation. Frost forms when lenses or windows are coated with ice crystals. Fog shrouds objects and is rendered with the diamond-square algorithm. Brightness varies with daylight intensity. Contrast can be high or low depending on lighting conditions and the photographed object's color. Elastic transformations stretch or contract small image regions. Pixelation occurs when upsampling a lowresolution image. JPEG is a lossy image compression format which introduces compression artifacts.

The specific parameters for each corruption can be found in the project's github at the corruptions file: `https://github.com/dooleys/Robustness-Disparities-in-Commercial-Face-Detection/blob/main/code/imagenet_c_big/corruptions.py`.

B Metric Discussion

Our use of relative error is slightly adapted from ImageNet-C insomuch that in that paper, they were measuring top-1 error of classification systems. However, the concept is identical. Consequently, it is linguistically best for our measure to be *mean relative* corruption error, whereas ImageNet-C reports the *mean relative* corruption error. This symmantic difference is attributed to when we take our average versus when the ImageNet-C protocol does.

B.1 Response Example

Our main metric, $mrCE$ relies on the computing the number of detected faces, i.e.,length of a response, l_r, from an API service. We explicitly give an example of this here.

An example response from an API service is below has one face detected:

```
[
  {
    "face_rectangle": {
      "width": 601,
      "height": 859,
      "left": 222,
      "top": 218
    }
  }
]
```

An example response from an API service is below has two faces detected:

```
[
  {
    "face_rectangle": {
      "width": 601,
      "height": 859,
      "left": 222,
```

```
        "top": 218
      }
    },
    {
      "face_rectangle": {
        "width": 93,
        "height": 120,
        "left": 10,
        "top": 39
      }
    }
]
```


B.2 Metric Example

Let us consider an example. Assume we were testing the question of whether there is a difference in error rates for different eye colors: (Grey, Hazel, and Brown). Across all the corrupted data, we might see that rCE is 0.12, 0.21, and 0.23 for Grey, Hazel, and Brown respectively. Recall that the odds of an event with likelihood p is reported as $p/(1-p)$. So the odds of error for each eye color is 0.14, 0.27, and 0.30 respectively. Our logistic regression would be written as $\texttt{rCE} = \beta_0 + \beta_1\texttt{Hazel} + \beta_2\texttt{Brown}$, with `Hazel` and `Brown` being indicator variables. After fitting our regression, we might see that the estimated *odds* coefficients for the intercept is 0.14, for the variable `Hazel` is 1.93, and for the variable Brown is 2.14; and all the coefficients are significant. This makes sense because the odds of Grey is 0.14, the odds ratio between Hazel and Grey is 0.27/0.14 = 1.93, and the odds ratio between Brown and Grey is 0.30/0.14 = 2.14. The significance tells us that the odds (or probability) or error for Grey eyes is significantly different from the odds (or probability) of error for Brown and Hazel eyes. In this example, we can conclude that the odds of error is 2.14 higher for Brown eyes compared to Grey. Put another way, for every 1 error for a Grey eyed person, there would be roughly 2 errors for a Hazel or Brown person. More so, the odds of error for Brown eyes is 114% higher than the odds of error for for Grey eyes.

C API Parameteres

For the AWS DetectFaces API,[2] we selected to have all facial attributes returned. This includes age and gender estimates. We evaluate the performance of these estimates in Section 5. The Azure Face API[3] allows the user to select one of three detection models. We chose model `detection_03` as it was their most recently released model (February 2021) and was described to have the highest performance on small, side, and blurry faces, since it aligns with our benchmark intention. This model does not return age or gender estimates (though model `detection_01` does).

D Benchmarks Costs

A total breakdown of costs for this benchmark can be found in Table 5.

E Statistical Significance Regressions for rCE

E.1 Service Comparison Claims

Regressions for $mrCE$ by service can be found in Table 6.

Regressions regarding the service difference for each dataset dataset can be found in Table 7

E.2 Corruption Comparison Claims

Regressions for the interaction of corruptions and service can be found in Table 8.

Regressions for the interaction of corruptions and service by dataset can be found in Table 9.

Regression of $mrCE$ by dataset and service can be found in Table 10.

[2] `https://docs.aws.amazon.com/rekognition/latest/dg/API_DetectFaces.html`

[3] `https://westus.dev.cognitive.microsoft.com/docs/services/563879b61984550e40cbbe8d/operations/563879b61984550f30395236`

E.3 Age Comparison Claims

Regression of $mrCE$ for Age for the Adience dataset can be found in Table 11.

Regression of $mrCE$ for Age for the CCD dataset can be found in Table 12.

Regression of $mrCE$ for Age as a numeric variable for the CCD dataset can be found in Table 13.

Regression of $mrCE$ for Age for the MIAP dataset can be found in Table 14.

Regression of $mrCE$ for Age as a categorical variable for the UTK dataset can be found in Table 15.

Regression of $mrCE$ for Age as a numeric variable for the UTK dataset can be found in Table 16.

Regression of $mrCE$ for Age interaction with corruption for the Adience dataset can be found in Table 17.

Regression of $mrCE$ for Age interaction with corruption for the CCD dataset can be found in Table 18.

Regression of $mrCE$ for Age interaction with corruption for the MIAP dataset can be found in Table 19.

Regression of $mrCE$ for Age interaction with corruption for the UTKFace dataset can be found in Table 20.

E.4 Gender Comparison Claims

Regression of $mrCE$ for Gender for the Adience dataset can be found in Table 21.

Regression of $mrCE$ for Gender for the CCD dataset can be found in Table 22.

Regression of $mrCE$ for Gender for the MIAP dataset can be found in Table 23.

Regression of $mrCE$ for Gender for the UTKFace dataset can be found in Table 24.

Regression of $mrCE$ for Gender interaction with corruption for the Adience dataset can be found in Table 25.

Regression of $mrCE$ for Gender interaction with corruption for the CCD dataset can be found in Table 26.

Regression of $mrCE$ for Gender interaction with corruption for the MIAP dataset can be found in Table 27.

Regression of $mrCE$ for Gender interaction with corruption for the UTKFace dataset can be found in Table 28.

E.5 Skin Type Comparison Claims

Regression of $mrCE$ for Skin Type for the CCD dataset can be found in Table 29.

Regression of $mrCE$ for Ethnicity for the UTKFace dataset can be found in Table 30.

Regression of $mrCE$ for Skin Type interaction with corruption for the CCD dataset can be found in Table 31.

Regression of $mrCE$ for Ethnicity interaction with corruption for the UTKFace dataset can be found in Table 32.

E.6 Lighting Comparison Claims

Regression of $mrCE$ for Lighting for the CCD dataset can be found in Table 33.

Regression of $mrCE$ for Lighting interaction with corruption for the CCD dataset can be found in Table 34.

E.7 Interactions Comparison Claims

E.7.1 CCD Demographic Interactions

Regression of $mrCE$ for the interaction of Lighting and Skin Type for the CCD dataset can be found in Table 35.

Regression of $mrCE$ for the interaction of Lighting and Age for the CCD dataset can be found in Table 36.

Regression of $mrCE$ for the interaction of Lighting and Age (as a numeric variable) for the CCD dataset can be found in Table 37.

Regression of $mrCE$ for the interaction of Lighting and Gender for the CCD dataset can be found in Table 38.

Regression of $mrCE$ for the interaction of Age and Gender for the CCD dataset can be found in Table 39.

Regression of $mrCE$ for the interaction of Age, Skin Type, and Gender for the CCD dataset can be found in Table 40.

Regression of $mrCE$ for the interaction of Age (as a numeric variable) and Gender for the CCD dataset can be found in Table 41.

Regression of $mrCE$ for the interaction of Age and Skin Type for the CCD dataset can be found in Table 42.

Regression of $mrCE$ for the interaction of Age (as a numeric variable) and Skin Type for the CCD dataset can be found in Table 43.

Regression of $mrCE$ for the interaction of Gender and Skin Type for the CCD dataset can be found in Table 44.

E.7.2 Adience Demographic Interactions

Regression of $mrCE$ for the interaction of Age and Gender for the Adience dataset can be found in Table 45.

E.7.3 MIAP Demographic Interactions

Regression of $mrCE$ for the interaction of Age and Gender for the MIAP dataset can be found in Table 46.

E.7.4 UTKFace Demographic Interactions

Regression of $mrCE$ for the interaction of Age and Gender for the UTKFace dataset can be found in Table 47.

Regression of $mrCE$ for the interaction of Age and Ethnicity for the UTKFace dataset can be found in Table 48.

Regression of $mrCE$ for the interaction of Gender and Ethnicity for the UTKFace dataset can be found in Table 49.

Regression of $mrCE$ for the interaction of Gender and Ethnicity for the UTKFace dataset can be found in Table 49.

F Statistical Significance Regressions for Gender Prediction

Regression of $mrCE$ on Gender Prediction by dataset can be found in Table 50.

F.1 Age Comparison Claims

Regressions for Gender Prediction for Age for the Adience dataset can be found in Table 51.

Regressions for Gender Prediction for for Age as a categorical variable for the CCD dataset can be found in Table 52.

Regressions for Gender Prediction for for Age as a numeric variable for the CCD dataset can be found in Table 53.

Regressions for Gender Prediction for Age for the MIAP dataset can be found in Table 54.

Regressions for Gender Prediction for Age for the UTKFace dataset can be found in Table 55.

Regressions for Gender Prediction for Age (as a numeric variable) for the UTKFace dataset can be found in Table 56.

F.2 Gender Comparison Claims

Regressions for Gender Prediction for Gender for the Adience dataset can be found in Table 57.

Regressions for Gender Prediction by Gender for the CCD dataset can be found in Table 58.

Regressions for Gender Prediction by Gender for the MIAP dataset can be found in Table 59.

Regressions for Gender Prediction by Gender for the UTKFace dataset can be found in Table 60.

F.3 Skin Type Comparison Claims

Regressions for Gender Prediction by Skin Type for the CCD dataset can be found in Table 61.

Regressions for Gender Prediction by Ethnicity for the UTKFace dataset can be found in Table 62.

F.4 Lighting Comparison Claims

Regressions for Gender Prediction by Lighting for the CCD dataset can be found in Table 63.

G Statistical Significance Regressions for Age Estimation

Regression of $mrCE$ on Age Estimation by dataset can be found in Table 64.

Regressions for Age Estimation for Age as a numeric variable for the CCD dataset can be found in Table 65.

Regressions for Age Estimation for Age (as a numeric variable) for the UTKFace dataset can be found in Table 66.

Regressions for Age Estimation by Gender for the CCD dataset can be found in Table 67.

Regressions for Age Estimation by Gender for the UTKFace dataset can be found in Table 68.

Regressions for Age Estimation by Skin Type for the CCD dataset can be found in Table 69.

Regressions for Age Estimation by Ethnicity for the UTKFace dataset can be found in Table 70.

Regressions for Age Estimation by Lighting for the CCD dataset can be found in Table 71.

List of Tables

Table 1: Adience Dataset Counts

Age	Gender	Count
0-2	Female	684
	Male	716
3-7	Female	1232
	Male	925
8-14	Female	1353
	Male	933
15-24	Female	1047
	Male	742
25-35	Female	1500
	Male	1500
36-45	Female	1078
	Male	1412
46-59	Female	436
	Male	466
60+	Female	428
	Male	467

Table 2: CCD Dataset Counts

Lighting	Gender	Skin	Age	Count
Bright	Female	Dark	19-45	1500
			45-64	1500
			65+	547
		Light	19-45	1500
			45-64	1500
			65+	653
	Male	Dark	19-45	1500
			45-64	1500
			65+	384
		Light	19-45	1500
			45-64	1500
			65+	695
	Other	Dark	19-45	368
			45-64	168
			65+	12
		Light	19-45	244
			45-64	49
Dim	Female	Dark	19-45	1500
			45-64	670
			65+	100
		Light	19-45	642
			45-64	314
			65+	131
	Male	Dark	19-45	1500
			45-64	387
			65+	48
		Light	19-45	485
			45-64	299
			65+	123
	Other	Dark	19-45	57
			45-64	26
			65+	3
		Light	19-45	27
			45-64	12

Table 3: MIAP Dataset Counts

AgePresentation	GenderPresentation	Count
Young	Unknown	1500
Middle	Predominantly Feminine	1500
	Predominantly Masculine	1500
	Unknown	561
Older	Predominantly Feminine	209
	Predominantly Masculine	748
	Unknown	24
Unknown	Predominantly Feminine	250
	Predominantly Masculine	402
	Unknown	1500

Table 4: UTKFace Dataset Counts

Age	Gender	Race	Count
0-18	Female	Asian	555
		Black	161
		Indian	350
		Others	338
		White	987
	Male	Asian	586
		Black	129
		Indian	277
		Others	189
		White	955
19-45	Female	Asian	1273
		Black	1500
		Indian	1203
		Others	575
		White	1500
	Male	Asian	730
		Black	1499
		Indian	1264
		Others	477
		White	1500
45-64	Female	Asian	39
		Black	206
		Indian	146
		Others	22
		White	802
	Male	Asian	180
		Black	401
		Indian	653
		Others	97
		White	1500
65+	Female	Asian	75
		Black	78
		Indian	43
		Others	10
		White	712
	Male	Asian	148
		Black	166
		Indian	91
		Others	5
		White	682

Table 5: Total Costs of Benchmark

Category	Cost
Azure Face Service	$4,270.58
AWS Rekognition	$4,270.66
S3	$701.99
EC2	$387.70
Tax	$256.24
Total	$9,887.17

Table 6: Service comparisons (odds ratio) for all data.

	Dependent variable:
	rCE
	Corrupted
serviceazure	1.004
	t = 1.929*
Constant	0.140
	t = −1,442.579***
Observations	9,998,278
Log Likelihood	−3,731,564.000
Akaike Inf. Crit.	7,463,133.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 7: Service comparision (odds ratio) for each dataset.

	Dependent variable:			
	rCE			
	Adience	CCD	MIAP	UTKFace
	(1)	(2)	(3)	(4)
serviceazure	0.657	1.676	0.845	1.168
	t = -122.655^{***}	t = 130.450^{***}	t = -34.861^{***}	t = 40.284^{***}
Constant	0.292	0.076	0.218	0.091
	t = -545.107^{***}	t = -837.391^{***}	t = -457.743^{***}	t = -853.298^{***}
Observations	2,237,753	3,216,600	1,228,938	3,314,987
Log Likelihood	−1,090,783.000	−978,792.400	−554,258.500	−1,001,254.000
Akaike Inf. Crit.	2,181,571.000	1,957,589.000	1,108,521.000	2,002,511.000
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 8: Corruption comparison (odds ratios) with interaction of the service.

	Dependent variable:		
	rCE		
	Interactions	Just AWS	Just Azure
	(1)	(2)	(3)
corruptiongaussian-noise	0.264 t = −312.385***	0.264 t = −312.385***	0.483 t = −196.734***
corruptionshot-noise	0.262 t = −313.871***	0.262 t = −313.871***	0.473 t = −201.890***
corruptionimpulse-noise	0.251 t = −319.791***	0.251 t = −319.791***	0.406 t = −235.967***
corruptiondefocus-blur	0.075 t = −381.020***	0.075 t = −381.020***	0.055 t = −372.263***
corruptionglass-blur	0.064 t = −377.547***	0.064 t = −377.547***	0.046 t = −363.980***
corruptionmotion-blur	0.075 t = −381.012***	0.075 t = −381.012***	0.067 t = −378.723***
corruptionzoom-blur	0.365 t = −257.382***	0.365 t = −257.382***	0.271 t = −308.492***
corruptionsnow	0.137 t = −373.698***	0.137 t = −373.698***	0.116 t = −379.596***
corruptionfrost	0.225 t = −333.258***	0.225 t = −333.258***	0.173 t = −359.134***
corruptionfog	0.065 t = −377.853***	0.065 t = −377.853***	0.048 t = −366.168***
corruptionbrightness	0.028 t = −335.288***	0.028 t = −335.288***	0.021 t = −316.901***
corruptioncontrast	0.134 t = −374.632***	0.134 t = −374.632***	0.104 t = −381.640***
corruptionelastic-transform	0.090 t = −382.614***	0.090 t = −382.614***	0.058 t = −374.277***
corruptionpixelate	0.083 t = −382.242***	0.083 t = −382.242***	0.035 t = −350.249***
corruptionjpeg-compression	0.099 t = −382.239***	0.099 t = −382.239***	0.044 t = −362.659***
serviceazure	1.829 t = 107.019***		
corruptionshot-noise:serviceazure	0.989 t = −1.447		
corruptionimpulse-noise:serviceazure	0.885 t = −15.102***		
corruptiondefocus-blur:serviceazure	0.400 t = −77.782***		
corruptionglass-blur:serviceazure	0.386 t = −76.060***		
corruptionmotion-blur:serviceazure	0.491 t = −62.806***		
corruptionzoom-blur:serviceazure	0.406 t = −111.717***		
corruptionsnow:serviceazure	0.463 t = −80.110***		
corruptionfrost:serviceazure	0.420 t = −99.791***		
corruptionfog:serviceazure	0.400 t = −74.066***		
corruptionbrightness:serviceazure	0.417 t = −51.002***		
corruptioncontrast:serviceazure	0.424 t = −87.596***		
corruptionelastic-transform:serviceazure	0.353 t = −91.725***		
corruptionpixelate:serviceazure	0.232 t = −113.499***		
corruptionjpeg-compression:serviceazure	0.245 t = −118.081***		
Observations	9,998,278	4,999,500	4,998,778
Log Likelihood	−3,381,198.000	−1,753,589.000	−1,627,609.000
Akaike Inf. Crit.	6,762,457.000	3,507,208.000	3,255,249.000
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 9: Corruption comparison (odds ratios) with interactions of the service by dataset.

	Dependent variable:			
	rCE			
	Adience	CCD	MIAP	UTKFace
	(1)	(2)	(3)	(4)
corruptiongaussian-noise	0.867	0.044	0.290	0.239
	t = −19.371***	t = −205.515***	t = −104.521***	t = −187.890***
corruptionshot-noise	0.867	0.019	0.282	0.269
	t = −19.451***	t = −175.607***	t = −106.080***	t = −178.325***
corruptionimpulse-noise	0.815	0.047	0.299	0.208
	t = −27.738***	t = −207.339***	t = −102.783***	t = −197.058***
corruptiondefocus-blur	0.129	0.008	0.181	0.075
	t = −177.875***	t = −139.187***	t = −124.688***	t = −219.327***
corruptionglass-blur	0.130	0.003	0.094	0.076
	t = −177.863***	t = −106.569***	t = −134.130***	t = −219.465***
corruptionmotion-blur	0.110	0.007	0.196	0.083
	t = −180.127***	t = −136.128***	t = −122.181***	t = −220.095***
corruptionzoom-blur	0.159	1.116	0.694	0.053
	t = −172.809***	t = 17.985***	t = −36.415***	t = −213.709***
corruptionsnow	0.295	0.118	0.245	0.035
	t = −139.967***	t = −214.967***	t = −113.144***	t = −200.981***
corruptionfrost	0.603	0.169	0.262	0.092
	t = −67.027***	t = −204.765***	t = −110.003***	t = −220.319***
corruptionfog	0.187	0.015	0.120	0.025
	t = −166.804***	t = −164.639***	t = −132.732***	t = −189.501***
corruptionbrightness	0.053	0.005	0.079	0.016
	t = −175.270***	t = −118.824***	t = −133.824***	t = −171.880***
corruptioncontrast	0.249	0.096	0.266	0.062
	t = −151.717***	t = −216.916***	t = −109.117***	t = −216.674***
corruptionelastic-transform	0.185	0.006	0.124	0.108
	t = −167.302***	t = −130.036***	t = −132.381***	t = −219.452***
corruptionpixelate	0.129	0.040	0.228	0.050
	t = −177.943***	t = −202.840***	t = −116.398***	t = −212.167***
corruptionjpeg-compression	0.354	0.008	0.120	0.050
	t = −124.648***	t = −141.366***	t = −132.780***	t = −212.281***
serviceazure	1.030	9.359	1.134	1.706
	t = 2.828***	t = 134.594***	t = 7.640***	t = 52.843***
corruptionshot-noise:serviceazure	1.098	1.526	1.017	1.100
	t = 6.401***	t = 14.583***	t = 0.714	t = 6.765***
corruptionimpulse-noise:serviceazure	0.846	0.740	0.941	1.000
	t = −11.357***	t = −12.909***	t = −2.617***	t = −0.026
corruptiondefocus-blur:serviceazure	0.389	0.453	0.647	0.411
	t = −41.182***	t = −18.688***	t = −16.508***	t = −42.659***
corruptionglass-blur:serviceazure	0.444	0.396	0.520	0.517
	t = −36.526***	t = −14.728***	t = −19.950***	t = −33.197***
corruptionmotion-blur:serviceazure	0.533	0.330	0.653	0.650
	t = −27.908***	t = −24.648***	t = −16.435***	t = −23.084***
corruptionzoom-blur:serviceazure	0.419	0.074	0.701	0.542
	t = −41.256***	t = −139.197***	t = −16.272***	t = −27.645***
corruptionsnow:serviceazure	0.598	0.110	0.783	0.655
	t = −30.377***	t = −101.571***	t = −10.062***	t = −16.793***
corruptionfrost:serviceazure	0.451	0.137	0.787	0.366
	t = −51.076***	t = −97.439***	t = −9.941***	t = −50.463***
corruptionfog:serviceazure	0.508	0.152	0.690	0.613
	t = −35.057***	t = −50.274***	t = −12.844***	t = −16.862***
corruptionbrightness:serviceazure	0.422	0.227	0.667	0.625
	t = −27.257***	t = −25.777***	t = −12.276***	t = −13.475***
corruptioncontrast:serviceazure	0.579	0.091	0.740	0.552
	t = −31.026***	t = −104.456***	t = −12.489***	t = −28.303***
corruptionelastic-transform:serviceazure	0.445	0.269	0.598	0.406
	t = −40.734***	t = −26.628***	t = −17.532***	t = −48.530***
corruptionpixelate:serviceazure	0.270	0.050	0.415	0.307
	t = −51.699***	t = −92.853***	t = −32.956***	t = −45.817***
corruptionjpeg-compression:serviceazure	0.228	0.251	0.630	0.344
	t = −77.267***	t = −31.449***	t = −15.725***	t = −42.722***
Observations	2,237,753	3,216,600	1,228,938	3,314,987
Log Likelihood	−937,818.300	−732,058.100	−524,575.100	−885,876.800
Akaike Inf. Crit.	1,875,697.000	1,464,176.000	1,049,210.000	1,771,814.000
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 10: Corruption comparison (odds ratios) by the service and by the dataset.

	Dependent variable:							
	rCE							
	Adience+AWS	Adience+Azure	CCD+AWS	CCD+Azure	MIAP+AWS	MIAP+Azure	UTKFace+AWS	UTKFace+Azure
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
corruptiongaussian-noise	0.867	0.893	0.044	0.412	0.290	0.329	0.239	0.407
	t = −19.371***	t = −15.388***	t = −205.515***	t = −131.997***	t = −104.521***	t = −97.041***	t = −187.890***	t = −135.493***
corruptionshot-noise	0.867	0.981	0.019	0.272	0.282	0.325	0.269	0.505
	t = −19.451***	t = −2.669***	t = −175.607***	t = −174.868***	t = −106.080***	t = −97.772***	t = −178.325***	t = −107.188***
corruptionimpulse-noise	0.815	0.710	0.047	0.326	0.299	0.320	0.208	0.355
	t = −27.738***	t = −46.051***	t = −207.339***	t = −158.077***	t = −102.783***	t = −98.916***	t = −197.058***	t = −151.411***
corruptiondefocus-blur	0.129	0.052	0.008	0.033	0.181	0.133	0.075	0.052
	t = −177.875***	t = −174.976***	t = −139.187***	t = −196.428***	t = −124.688***	t = −131.472***	t = −219.327***	t = −213.209***
corruptionglass-blur	0.130	0.059	0.003	0.012	0.094	0.055	0.076	0.067
	t = −177.863***	t = −177.368***	t = −106.569***	t = −156.851***	t = −134.130***	t = −130.543***	t = −219.465***	t = −217.929***
corruptionmotion-blur	0.110	0.060	0.007	0.022	0.196	0.145	0.083	0.092
	t = −180.127***	t = −177.676***	t = −136.128***	t = −181.776***	t = −122.181***	t = −129.983***	t = −220.095***	t = −220.287***
corruptionzoom-blur	0.159	0.069	1.116	0.769	0.694	0.551	0.053	0.049
	t = −172.809***	t = −179.368***	t = 17.985***	t = −42.717***	t = −36.415***	t = −57.733***	t = −213.709***	t = −211.771***
corruptionsnow	0.295	0.181	0.118	0.122	0.245	0.218	0.035	0.039
	t = −139.967***	t = −168.115***	t = −214.967***	t = −214.413***	t = −113.144***	t = −118.231***	t = −200.981***	t = −204.610***
corruptionfrost	0.603	0.280	0.169	0.217	0.262	0.234	0.092	0.057
	t = −67.027***	t = −143.805***	t = −204.765***	t = −191.509***	t = −110.003***	t = −115.280***	t = −220.319***	t = −215.136***
corruptionfog	0.187	0.098	0.015	0.021	0.120	0.094	0.025	0.026
	t = −166.804***	t = −180.880***	t = −164.639***	t = −178.835***	t = −132.732***	t = −134.110***	t = −189.501***	t = −191.213***
corruptionbrightness	0.053	0.023	0.005	0.010	0.079	0.060	0.016	0.017
	t = −175.279***	t = −152.540***	t = −118.824***	t = −148.220***	t = −133.824***	t = −131.498***	t = −171.880***	t = −174.501***
corruptioncontrast	0.249	0.148	0.096	0.082	0.266	0.224	0.062	0.058
	t = −151.717***	t = −174.691***	t = −216.916***	t = −216.753***	t = −109.117***	t = −117.161***	t = −216.674***	t = −215.481***
corruptionelastic-transform	0.185	0.085	0.006	0.016	0.124	0.084	0.108	0.075
	t = −167.302***	t = −180.901***	t = −130.036***	t = −167.391***	t = −132.381***	t = −134.031***	t = −219.452***	t = −219.315***
corruptionpixelate	0.129	0.036	0.040	0.019	0.228	0.107	0.050	0.026
	t = −177.943***	t = −166.180***	t = −202.840***	t = −174.946***	t = −116.398***	t = −133.643***	t = −212.167***	t = −191.075***
corruptionjpeg-compression	0.354	0.083	0.008	0.019	0.120	0.085	0.050	0.030
	t = −124.648***	t = −180.834***	t = −141.366***	t = −176.195***	t = −132.780***	t = −134.068***	t = −212.281***	t = −195.447***
Observations	1,118,925	1,118,828	1,608,300	1,608,300	614,550	614,388	1,657,725	1,657,262
Log Likelihood	−537,021.700	−400,796.600	−284,403.200	−447,655.000	−275,816.000	−248,759.000	−442,257.100	−443,619.600
Akaike Inf. Crit.	1,074,073.000	801,623.200	568,836.300	895,339.900	551,662.100	497,548.000	884,544.300	887,269.300

Note: *p<0.1; **p<0.05; ***p<0.01

Table 11: Age comparison (odds ratios) for the Adience dataset with each age group a reference label.

	Dependent variable:							
	rCE							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
age_group0-2		1.099 t = 12.775***	1.050 t = 6.680***	0.958 t = −5.692***	0.770 t = −38.721***	0.828 t = −26.947***	0.857 t = −17.490***	0.772 t = −29.643***
age_group3-7	0.910 t = −12.775***		0.955 t = −7.022***	0.871 t = −20.053***	0.701 t = −59.789***	0.754 t = −45.460***	0.779 t = −30.194***	0.703 t = −43.483***
age_group8-14	0.953 t = −6.680***	1.047 t = 7.022***		0.912 t = −13.628***	0.734 t = −53.508***	0.789 t = −39.031***	0.816 t = −24.972***	0.736 t = −38.370***
age_group15-24	1.044 t = 5.692***	1.148 t = 20.053***	1.096 t = 13.628***		0.804 t = −35.504***	0.865 t = −22.660***	0.894 t = −13.300***	0.806 t = −26.069***
age_group25-35	1.298 t = 38.721***	1.427 t = 59.789***	1.363 t = 53.508***	1.244 t = 35.504***		1.076 t = 13.477***	1.112 t = 13.860***	1.003 t = 0.353
age_group36-45	1.207 t = 26.947***	1.327 t = 45.460***	1.267 t = 39.031***	1.156 t = 22.660***	0.930 t = −13.477***		1.034 t = 4.238***	0.932 t = −9.082***
age_group46-59	1.167 t = 17.490***	1.283 t = 30.194***	1.226 t = 24.972***	1.118 t = 13.300***	0.899 t = −13.860***	0.967 t = −4.238***		0.902 t = −10.963***
age_group60+	1.295 t = 29.643***	1.423 t = 43.483***	1.359 t = 38.370***	1.240 t = 26.069***	0.997 t = −0.353	1.073 t = 9.082***	1.109 t = 10.963***	
Constant	0.217 t = −267.982***	0.197 t = −342.431***	0.207 t = −347.774***	0.227 t = −298.458***	0.282 t = −351.876***	0.262 t = −332.013***	0.253 t = −202.799***	0.281 t = −192.452***
Observations	2,237,753	2,237,753	2,237,753	2,237,753	2,237,753	2,237,753	2,237,753	2,237,753
Log Likelihood	−1,095,217.000	−1,095,217.000	−1,095,217.000	−1,095,217.000	−1,095,217.000	−1,095,217.000	−1,095,217.000	−1,095,217.000
Akaike Inf. Crit.	2,190,450.000	2,190,450.000	2,190,450.000	2,190,450.000	2,190,450.000	2,190,450.000	2,190,450.000	2,190,450.000

Note: *p<0.1; **p<0.05; ***p<0.01

Table 12: Age comparison (odds ratios) for the CCD dataset with each age group a reference label.

	Dependent variable:		
	rCE		
	(1)	(2)	(3)
Age19-45		0.958	0.900
		t = -10.352^{***}	t = -17.691^{***}
Age45-64	1.044		0.940
	t = 10.352^{***}		t = -10.074^{***}
Age65+	1.111	1.064	
	t = 17.691^{***}	t = 10.074^{***}	
Constant	0.098	0.103	0.109
	t = -843.752^{***}	t = -721.180^{***}	t = -419.564^{***}
Observations	3,216,600	3,216,600	3,216,600
Log Likelihood	$-987,347.400$	$-987,347.400$	$-987,347.400$
Akaike Inf. Crit.	1,974,701.000	1,974,701.000	1,974,701.000
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 13: Age comparison for the CCD dataset with each age as a numeric value.

	Dependent variable:
	rCE
Age_Numeric	0.0002^{***}
	(0.00001)
Constant	0.082^{***}
	(0.0004)
Observations	3,216,600
Log Likelihood	−570,315.400
Akaike Inf. Crit.	1,140,635.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 14: Age comparison (odds ratios) for the MIAP dataset with each age group a reference label.

	Dependent variable:			
	rCE			
	(1)	(2)	(3)	(4)
AgePresentationYoung		1.237 t = 24.175***	0.565 t = −101.148***	1.077 t = 10.365***
AgePresentationMiddle	0.808 t = −24.175***		0.456 t = −87.797***	0.871 t = −13.840***
AgePresentationOlder	1.771 t = 101.148***	2.191 t = 87.797***		1.908 t = 87.814***
AgePresentationUnknown	0.928 t = −10.365***	1.148 t = 13.840***	0.524 t = −87.814***	
Constant	0.175 t = −453.546***	0.141 t = −247.141***	0.310 t = −282.848***	0.162 t = −298.918***
Observations	1,228,938	1,228,938	1,228,938	1,228,938
Log Likelihood	−547,430.100	−547,430.100	−547,430.100	−547,430.100
Akaike Inf. Crit.	1,094,868.000	1,094,868.000	1,094,868.000	1,094,868.000
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 15: Age comparison (odds ratios) for the UTK dataset with each age group a reference label.

	Dependent variable:			
	rCE			
	(1)	(2)	(3)	(4)
Age0-18		1.134 t = 25.101***	0.986 t = −2.377**	0.794 t = −32.553***
Age19-45	0.882 t = −25.101***		0.869 t = −27.005***	0.700 t = −56.082***
Age45-64	1.015 t = 2.377**	1.150 t = 27.005***		0.805 t = −29.975***
Age65+	1.260 t = 32.553***	1.428 t = 56.082***	1.242 t = 29.975***	
Constant	0.103 t = −544.932***	0.090 t = −871.235***	0.104 t = −515.001***	0.129 t = −357.390***
Observations	3,314,537	3,314,537	3,314,537	3,314,537
Log Likelihood	−1,000,264.000	−1,000,264.000	−1,000,264.000	−1,000,264.000
Akaike Inf. Crit.	2,000,535.000	2,000,535.000	2,000,535.000	2,000,535.000
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 16: Age comparison for the UTK dataset with each age as a numeric value.

	Dependent variable:
	rCE
Age_Numeric	0.0003^{***}
	(0.00001)
Constant	0.081^{***}
	(0.0003)
Observations	3,314,987
Log Likelihood	−553,811.000
Akaike Inf. Crit.	1,107,626.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 17: Age comparison (odds ratios) with interaction with corruption for the Adience dataset.

	Dependent variable:
	rCE
age_group3-7	0.889 t = −5.393***
age_group8-14	0.883 t = −5.789***
age_group15-24	0.919 t = −3.754***
age_group25-35	1.062 t = 2.944***
age_group36-45	0.996 t = −0.196
age_group46-59	0.960 t = −1.508
age_group60+	1.131 t = 4.535***
corruptionshot-noise	1.162 t = 6.286***
corruptionimpulse-noise	0.881 t = −5.299***
corruptiondefocus-blur	0.070 t = −67.356***
corruptionglass-blur	0.055 t = −67.174***
corruptionmotion-blur	0.053 t = −67.048***
corruptionzoom-blur	0.095 t = −66.197***
corruptionsnow	0.185 t = −57.275***
corruptionfrost	0.470 t = −30.181***
corruptionfog	0.098 t = −65.978***
corruptionbrightness	0.022 t = −60.357***
corruptioncontrast	0.192 t = −56.493***
corruptionelastic-transform	0.101 t = −65.724***
corruptionpixelate	0.080 t = −67.049***
corruptionjpeg-compression	0.219 t = −53.549***
age_group3-7:corruptionshot-noise	0.929 t = −2.397**
age_group8-14:corruptionshot-noise	0.896 t = −3.594***
age_group15-24:corruptionshot-noise	0.851 t = −5.046***
age_group25-35:corruptionshot-noise	0.907 t = −3.368***
age_group36-45:corruptionshot-noise	0.878 t = −4.356***
age_group46-59:corruptionshot-noise	0.850 t = −4.245***
age_group60+:corruptionshot-noise	0.905 t = −2.595***
age_group3-7:corruptionimpulse-noise	0.995 t = −0.159
age_group8-14:corruptionimpulse-noise	0.965 t = −1.173
age_group15-24:corruptionimpulse-noise	0.933 t = −2.164**
age_group25-35:corruptionimpulse-noise	0.996 t = −0.125
age_group36-45:corruptionimpulse-noise	0.979 t = −0.712
age_group46-59:corruptionimpulse-noise	0.986 t = −0.368
age_group60+:corruptionimpulse-noise	1.019 t = 0.494
age_group3-7:corruptiondefocus-blur	0.885 t = −2.306**
age_group8-14:corruptiondefocus-blur	1.052 t = 0.989
age_group15-24:corruptiondefocus-blur	1.585 t = 9.177***
age_group25-35:corruptiondefocus-blur	1.944 t = 14.774***
age_group36-45:corruptiondefocus-blur	1.838 t = 13.096***
age_group46-59:corruptiondefocus-blur	1.637 t = 8.505***
age_group60+:corruptiondefocus-blur	1.289 t = 4.320***
age_group3-7:corruptionglass-blur	1.006 t = 0.099
age_group8-14:corruptionglass-blur	1.359 t = 5.738***
age_group15-24:corruptionglass-blur	1.980 t = 12.875***
age_group25-35:corruptionglass-blur	2.904 t = 22.331***
age_group36-45:corruptionglass-blur	2.450 t = 18.162***
age_group46-59:corruptionglass-blur	2.146 t = 12.707***
age_group60+:corruptionglass-blur	1.669 t = 8.393***
age_group3-7:corruptionmotion-blur	0.965 t = −0.610
age_group8-14:corruptionmotion-blur	1.216 t = 3.545***
age_group15-24:corruptionmotion-blur	1.888 t = 11.720***
age_group25-35:corruptionmotion-blur	2.659 t = 20.067***
age_group36-45:corruptionmotion-blur	2.326 t = 16.777***
age_group46-59:corruptionmotion-blur	2.163 t = 12.653***
age_group60+:corruptionmotion-blur	1.775 t = 9.349***
age_group3-7:corruptionzoom-blur	0.788 t = −4.959***
age_group8-14:corruptionzoom-blur	0.946 t = −1.210
age_group15-24:corruptionzoom-blur	1.380 t = 6.999***
age_group25-35:corruptionzoom-blur	1.899 t = 15.723***
age_group36-45:corruptionzoom-blur	1.633 t = 11.566***
age_group46-59:corruptionzoom-blur	1.437 t = 6.763***
age_group60+:corruptionzoom-blur	1.190 t = 3.216***
age_group3-7:corruptionsnow	1.215 t = 5.184***
age_group8-14:corruptionsnow	1.508 t = 11.230***
age_group15-24:corruptionsnow	1.475 t = 10.159***
age_group25-35:corruptionsnow	1.432 t = 10.345***
age_group36-45:corruptionsnow	1.553 t = 12.345***
age_group46-59:corruptionsnow	1.773 t = 12.958***
age_group60+:corruptionsnow	1.868 t = 14.364***
age_group3-7:corruptionfrost	1.055 t = 1.673*
age_group8-14:corruptionfrost	1.058 t = 1.768*
age_group15-24:corruptionfrost	1.027 t = 0.787
age_group25-35:corruptionfrost	0.953 t = −1.604
age_group36-45:corruptionfrost	1.011 t = 0.351
age_group46-59:corruptionfrost	1.041 t = 1.009
age_group60+:corruptionfrost	1.145 t = 3.432***
age_group3-7:corruptionfog	1.440 t = 8.302***
age_group8-14:corruptionfog	1.813 t = 13.961***
age_group15-24:corruptionfog	1.649 t = 11.185***
age_group25-35:corruptionfog	1.642 t = 12.173***
age_group36-45:corruptionfog	1.784 t = 13.870***
age_group46-59:corruptionfog	1.925 t = 12.849***
age_group60+:corruptionfog	1.828 t = 11.969***
age_group3-7:corruptionbrightness	1.027 t = 0.323
age_group8-14:corruptionbrightness	1.526 t = 5.520***
age_group15-24:corruptionbrightness	1.953 t = 8.737***
age_group25-35:corruptionbrightness	2.604 t = 13.840***
age_group36-45:corruptionbrightness	2.648 t = 13.795***
age_group46-59:corruptionbrightness	2.353 t = 10.160***
age_group60+:corruptionbrightness	1.415 t = 3.850***
age_group3-7:corruptioncontrast	1.123 t = 3.086***
age_group8-14:corruptioncontrast	1.163 t = 4.083***
age_group15-24:corruptioncontrast	1.105 t = 2.562**
age_group25-35:corruptioncontrast	1.157 t = 4.196***
age_group36-45:corruptioncontrast	1.223 t = 5.609***
age_group46-59:corruptioncontrast	1.331 t = 6.323***
age_group60+:corruptioncontrast	1.236 t = 4.734***
age_group3-7:corruptionelastic-transform	0.986 t = −0.300
age_group8-14:corruptionelastic-transform	1.136 t = 2.881***
age_group15-24:corruptionelastic-transform	1.551 t = 9.850***
age_group25-35:corruptionelastic-transform	1.916 t = 16.269***
age_group36-45:corruptionelastic-transform	1.765 t = 13.746***
age_group46-59:corruptionelastic-transform	1.770 t = 11.184***
age_group60+:corruptionelastic-transform	1.574 t = 8.929***
age_group3-7:corruptionpixelate	0.699 t = −6.849***
age_group8-14:corruptionpixelate	0.759 t = −5.419***
age_group15-24:corruptionpixelate	1.121 t = 2.288**
age_group25-35:corruptionpixelate	1.546 t = 9.991***
age_group36-45:corruptionpixelate	1.421 t = 7.745***
age_group46-59:corruptionpixelate	1.289 t = 4.396***
age_group60+:corruptionpixelate	0.983 t = −0.286
age_group3-7:corruptionjpeg-compression	0.883 t = −3.342***
age_group8-14:corruptionjpeg-compression	0.900 t = −2.868***
age_group15-24:corruptionjpeg-compression	1.000 t = 0.010
age_group25-35:corruptionjpeg-compression	1.157 t = 4.296***
age_group36-45:corruptionjpeg-compression	1.155 t = 4.113***
age_group46-59:corruptionjpeg-compression	1.213 t = 4.353***
age_group60+:corruptionjpeg-compression	1.175 t = 3.676***
Constant	0.907 t = −5.779***
Observations	2,237,753
Log Likelihood	−947,029.700
Akaike Inf. Crit.	1,894,299.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 18: Age comparison (odds ratios) with interaction with corruption for the CCD dataset.

	Dependent variable:
	rCE
Age45-64	0.983
	t = −1.333
Age65+	1.031
	t = 1.673*
corruptionshot-noise	0.537
	t = −47.514***
corruptionimpulse-noise	0.854
	t = −13.361***
corruptiondefocus-blur	0.096
	t = −98.671***
corruptionglass-blur	0.036
	t = −89.949***
corruptionmotion-blur	0.068
	t = −97.482***
corruptionzoom-blur	4.130
	t = 139.345***
corruptionsnow	0.641
	t = −35.452***
corruptionfrost	1.031
	t = 2.670***
corruptionfog	0.098
	t = −98.663***
corruptionbrightness	0.032
	t = −88.094***
corruptioncontrast	0.413
	t = −63.079***
corruptionelastic-transform	0.054
	t = −95.275***
corruptionpixelate	0.135
	t = −97.037***
corruptionjpeg-compression	0.068
	t = −97.394***
Age45-64:corruptionshot-noise	1.405
	t = 17.577***
Age65+:corruptionshot-noise	1.614
	t = 17.830***
Age45-64:corruptionimpulse-noise	0.987
	t = −0.688
Age65+:corruptionimpulse-noise	0.989
	t = −0.418
Age45-64:corruptiondefocus-blur	1.104
	t = 2.750***
Age65+:corruptiondefocus-blur	1.126
	t = 2.350**
Age45-64:corruptionglass-blur	1.116
	t = 1.969**
Age65+:corruptionglass-blur	1.169
	t = 2.012**
Age45-64:corruptionmotion-blur	1.137
	t = 3.124***
Age65+:corruptionmotion-blur	1.137
	t = 2.211**
Age45-64:corruptionzoom-blur	1.220
	t = 12.705***
Age65+:corruptionzoom-blur	1.367
	t = 13.776***
Age45-64:corruptionsnow	0.885
	t = −6.180***
Age65+:corruptionsnow	0.837
	t = −6.176***
Age45-64:corruptionfrost	0.872
	t = −7.616***
Age65+:corruptionfrost	0.833
	t = −6.986***
Age45-64:corruptionfog	0.830
	t = −4.862***
Age65+:corruptionfog	0.688
	t = −6.271***
Age45-64:corruptionbrightness	1.318
	t = 4.900***
Age65+:corruptionbrightness	1.075
	t = 0.861
Age45-64:corruptioncontrast	1.146
	t = 6.428***
Age65+:corruptioncontrast	1.201
	t = 6.106***
Age45-64:corruptionelastic-transform	1.040
	t = 0.841
Age65+:corruptionelastic-transform	0.964
	t = −0.527
Age45-64:corruptionpixelate	1.172
	t = 5.167***
Age65+:corruptionpixelate	1.156
	t = 3.314***
Age45-64:corruptionjpeg-compression	1.073
	t = 1.673*
Age65+:corruptionjpeg-compression	0.896
	t = −1.731*
Constant	0.201
	t = −197.069***
Observations	3,216,600
Log Likelihood	−768,937.300
Akaike Inf. Crit.	1,537,965.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 19: Age comparison (odds ratios) with interaction with corruption for the MIAP dataset.

	Dependent variable:
	rCE
AgePresentationMiddle	0.830
	t = −6.436***
AgePresentationOlder	1.606
	t = 24.148***
AgePresentationUnknown	1.004
	t = 0.158
corruptionshot-noise	0.940
	t = −3.352***
corruptionimpulse-noise	1.001
	t = 0.036
corruptiondefocus-blur	0.472
	t = −35.769***
corruptionglass-blur	0.190
	t = −60.247***
corruptionmotion-blur	0.523
	t = −31.570***
corruptionzoom-blur	2.425
	t = 52.692***
corruptionsnow	0.694
	t = −18.852***
corruptionfrost	0.751
	t = −15.048***
corruptionfog	0.294
	t = −51.255***
corruptionbrightness	0.176
	t = −61.308***
corruptioncontrast	0.779
	t = −13.197***
corruptionelastic-transform	0.283
	t = −52.204***
corruptionpixelate	0.541
	t = −30.153***
corruptionjpeg-compression	0.282
	t = −52.289***
AgePresentationMiddle:corruptionshot-noise	1.113
	t = 2.632***
AgePresentationOlder:corruptionshot-noise	1.048
	t = 1.676*
AgePresentationUnknown:corruptionshot-noise	1.095
	t = 2.706***
AgePresentationMiddle:corruptionimpulse-noise	0.964
	t = −0.893
AgePresentationOlder:corruptionimpulse-noise	1.025
	t = 0.887
AgePresentationUnknown:corruptionimpulse-noise	0.970
	t = −0.906
AgePresentationMiddle:corruptiondefocus-blur	0.706
	t = −6.733***
AgePresentationOlder:corruptiondefocus-blur	1.402
	t = 11.098***
AgePresentationUnknown:corruptiondefocus-blur	0.793
	t = −5.763***
AgePresentationMiddle:corruptionglass-blur	0.624
	t = −6.412***
AgePresentationOlder:corruptionglass-blur	2.007
	t = 18.935***
AgePresentationUnknown:corruptionglass-blur	0.674
	t = −7.014***
AgePresentationMiddle:corruptionmotion-blur	0.856
	t = −3.243***
AgePresentationOlder:corruptionmotion-blur	1.283
	t = 8.282***
AgePresentationUnknown:corruptionmotion-blur	0.802
	t = −5.646***
AgePresentationMiddle:corruptionzoom-blur	1.077
	t = 1.989**
AgePresentationOlder:corruptionzoom-blur	0.554
	t = −22.326***
AgePresentationUnknown:corruptionzoom-blur	0.833
	t = −5.894***
AgePresentationMiddle:corruptionsnow	1.113
	t = 2.495**
AgePresentationOlder:corruptionsnow	1.187
	t = 5.944***
AgePresentationUnknown:corruptionsnow	1.016
	t = 0.443
AgePresentationMiddle:corruptionfrost	1.174
	t = 3.824***
AgePresentationOlder:corruptionfrost	1.070
	t = 2.346**
AgePresentationUnknown:corruptionfrost	1.141
	t = 3.833***
AgePresentationMiddle:corruptionfog	0.651
	t = −6.974***
AgePresentationOlder:corruptionfog	1.703
	t = 16.037***
AgePresentationUnknown:corruptionfog	0.813
	t = −4.496***
AgePresentationMiddle:corruptionbrightness	0.759
	t = −3.902***
AgePresentationOlder:corruptionbrightness	1.931
	t = 17.404***
AgePresentationUnknown:corruptionbrightness	0.764
	t = −4.814***
AgePresentationMiddle:corruptioncontrast	1.163
	t = 3.619***
AgePresentationOlder:corruptioncontrast	0.966
	t = −1.186
AgePresentationUnknown:corruptioncontrast	1.044
	t = 1.232
AgePresentationMiddle:corruptionelastic-transform	0.696
	t = −5.934***
AgePresentationOlder:corruptionelastic-transform	1.765
	t = 17.012***
AgePresentationUnknown:corruptionelastic-transform	0.713
	t = −7.055***
AgePresentationMiddle:corruptionpixelate	0.775
	t = −5.249***
AgePresentationOlder:corruptionpixelate	1.100
	t = 3.144***
AgePresentationUnknown:corruptionpixelate	0.799
	t = −5.775***
AgePresentationMiddle:corruptionjpeg-compression	0.692
	t = −6.022***
AgePresentationOlder:corruptionjpeg-compression	1.672
	t = 15.303***
AgePresentationUnknown:corruptionjpeg-compression	0.808
	t = −4.577***
Constant	0.276
	t = −100.098***
Observations	1,228,938
Log Likelihood	−515,664.800
Akaike Inf. Crit.	1,031,450.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 20: Age comparison (odds ratios) with interaction with corruption for the UTKFace dataset.

	Dependent variable:
	rCE
Age19-45	0.973
	t = −2.087**
Age45-64	1.226
	t = 12.883***
Age65+	1.437
	t = 19.046***
corruptionshot-noise	1.361
	t = 20.212***
corruptionimpulse-noise	0.939
	t = −3.951***
corruptiondefocus-blur	0.250
	t = −64.234***
corruptionglass-blur	0.281
	t = −60.959***
corruptionmotion-blur	0.404
	t = −48.142***
corruptionzoom-blur	0.190
	t = −70.171***
corruptionsnow	0.104
	t = −76.260***
corruptionfrost	0.250
	t = −64.234***
corruptionfog	0.074
	t = −76.140***
corruptionbrightness	0.044
	t = −72.689***
corruptioncontrast	0.260
	t = −63.135***
corruptionelastic-transform	0.273
	t = −61.818***
corruptionpixelate	0.115
	t = −75.861***
corruptionjpeg-compression	0.115
	t = −75.850***
Age19-45:corruptionshot-noise	0.810
	t = −11.639***
Age45-64:corruptionshot-noise	0.894
	t = −5.149***
Age65+:corruptionshot-noise	0.921
	t = −3.149***
Age19-45:corruptionimpulse-noise	0.902
	t = −5.452***
Age45-64:corruptionimpulse-noise	0.920
	t = −3.693***
Age65+:corruptionimpulse-noise	0.952
	t = −1.799*
Age19-45:corruptiondefocus-blur	0.830
	t = −7.135***
Age45-64:corruptiondefocus-blur	0.607
	t = −15.145***
Age65+:corruptiondefocus-blur	0.682
	t = −9.904***
Age19-45:corruptionglass-blur	0.829
	t = −7.463***
Age45-64:corruptionglass-blur	0.597
	t = −16.222***
Age65+:corruptionglass-blur	0.699
	t = −9.688***
Age19-45:corruptionmotion-blur	0.620
	t = −20.512***
Age45-64:corruptionmotion-blur	0.511
	t = −22.916***
Age65+:corruptionmotion-blur	0.712
	t = −10.216***
Age19-45:corruptionzoom-blur	0.893
	t = −3.983***
Age45-64:corruptionzoom-blur	0.685
	t = −10.695***
Age65+:corruptionzoom-blur	0.712
	t = −8.124***
Age19-45:corruptionsnow	1.201
	t = 5.315***
Age45-64:corruptionsnow	0.963
	t = −0.912
Age65+:corruptionsnow	1.116
	t = 2.325**
Age19-45:corruptionfrost	0.927
	t = −2.958***
Age45-64:corruptionfrost	0.876
	t = −4.279***
Age65+:corruptionfrost	0.946
	t = −1.548
Age19-45:corruptionfog	1.239
	t = 5.418***
Age45-64:corruptionfog	0.968
	t = −0.669
Age65+:corruptionfog	0.937
	t = −1.156
Age19-45:corruptionbrightness	1.454
	t = 7.671***
Age45-64:corruptionbrightness	0.988
	t = −0.206
Age65+:corruptionbrightness	0.868
	t = −1.966**
Age19-45:corruptioncontrast	0.643
	t = −16.680***
Age45-64:corruptioncontrast	0.640
	t = −13.865***
Age65+:corruptioncontrast	0.766
	t = −7.169***
Age19-45:corruptionelastic-transform	1.102
	t = 3.935***
Age45-64:corruptionelastic-transform	0.944
	t = −1.948*
Age65+:corruptionelastic-transform	1.113
	t = 3.127***
Age19-45:corruptionpixelate	1.189
	t = 5.216***
Age45-64:corruptionpixelate	0.840
	t = −4.240***
Age65+:corruptionpixelate	0.861
	t = −3.106***
Age19-45:corruptionjpeg-compression	1.112
	t = 3.183***
Age45-64:corruptionjpeg-compression	1.019
	t = 0.467
Age65+:corruptionjpeg-compression	1.242
	t = 4.883***
Constant	0.299
	t = −108.134***
Observations	3,314,537
Log Likelihood	−889,266.500
Akaike Inf. Crit.	1,778,653.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 21: Gender comparison (odds ratios) for the Adience dataset with each gender group a reference label.

	Dependent variable:	
	rCE	
	(1)	(2)
genderFemale		0.940
		t = -18.137^{***}
genderMale	1.063	
	t = 18.137^{***}	
Constant	0.232	0.247
	t = -615.718^{***}	t = -577.417^{***}
Observations	2,237,753	2,237,753
Log Likelihood	$-1,098,230.000$	$-1,098,230.000$
Akaike Inf. Crit.	2,196,465.000	2,196,465.000
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 22: Gender comparison (odds ratios) for the CCD dataset with each gender group a reference label.

	Dependent variable:		
	rCE		
	(1)	(2)	(3)
GenderFemale		0.920	1.007
		t = -21.077^{***}	t = 0.738
GenderMale	1.087		1.095
	t = 21.077^{***}		t = 9.335^{***}
GenderOther	0.993	0.914	
	t = -0.738	t = -9.335^{***}	
Constant	0.097	0.106	0.097
	t = -833.465^{***}	t = -805.996^{***}	t = -252.150^{***}
Observations	3,216,600	3,216,600	3,216,600
Log Likelihood	$-987{,}280.800$	$-987{,}280.800$	$-987{,}280.800$
Akaike Inf. Crit.	1,974,568.000	1,974,568.000	1,974,568.000
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 23: Gender comparison (odds ratios) for the MIAP dataset with each gender group a reference label.

	Dependent variable:		
	rCE		
	(1)	(2)	(3)
GenderPresentationPredominantly Feminine		0.870	0.664
		t = -20.124^{***}	t = -64.279^{***}
GenderPresentationPredominantly Masculine	1.150		0.763
	t = 20.124^{***}		t = -48.444^{***}
GenderPresentationUnknown	1.507	1.311	
	t = 64.279^{***}	t = 48.444^{***}	
Constant	0.159	0.183	0.239
	t = -342.987^{***}	t = -387.309^{***}	t = -413.812^{***}
Observations	1,228,938	1,228,938	1,228,938
Log Likelihood	$-552,371.200$	$-552,371.200$	$-552,371.200$
Akaike Inf. Crit.	1,104,748.000	1,104,748.000	1,104,748.000
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 24: Gender comparison (odds ratios) for the UTKFace dataset with each gender group a reference label.

	Dependent variable:	
	rCE	
	(1)	(2)
genderMale	1.004	
	t = 1.100	
genderFemale		0.996
		t = -1.100
Constant	0.099	0.099
	t = -833.743^{***}	t = -870.630^{***}
Observations	3,314,387	3,314,387
Log Likelihood	$-1,001,857.000$	$-1,001,857.000$
Akaike Inf. Crit.	2,003,718.000	2,003,718.000
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 25: Gender comparison (odds ratios) with interaction with corruption for the Adience dataset.

	Dependent variable:
	rCE
genderMale	1.105
	t = 9.595***
corruptionshot-noise	1.033
	t = 3.234***
corruptionimpulse-noise	0.853
	t = −15.495***
corruptiondefocus-blur	0.104
	t = −150.344***
corruptionglass-blur	0.110
	t = −149.224***
corruptionmotion-blur	0.098
	t = −151.416***
corruptionzoom-blur	0.136
	t = −144.013***
corruptionsnow	0.274
	t = −110.660***
corruptionfrost	0.495
	t = −65.801***
corruptionfog	0.171
	t = −135.614***
corruptionbrightness	0.044
	t = −151.179***
corruptioncontrast	0.225
	t = −122.456***
corruptionelastic-transform	0.155
	t = −139.503***
corruptionpixelate	0.092
	t = −152.182***
corruptionjpeg-compression	0.228
	t = −121.615***
genderMale:corruptionshot-noise	1.028
	t = 1.913*
genderMale:corruptionimpulse-noise	1.028
	t = 1.870*
genderMale:corruptiondefocus-blur	0.946
	t = −2.570**
genderMale:corruptionglass-blur	0.918
	t = −4.014***
genderMale:corruptionmotion-blur	0.970
	t = −1.374
genderMale:corruptionzoom-blur	0.874
	t = −6.672***
genderMale:corruptionsnow	0.948
	t = −3.167***
genderMale:corruptionfrost	0.940
	t = −4.047***
genderMale:corruptionfog	0.866
	t = −7.630***
genderMale:corruptionbrightness	0.935
	t = −2.262**
genderMale:corruptioncontrast	0.985
	t = −0.861
genderMale:corruptionelastic-transform	0.940
	t = −3.222***
genderMale:corruptionpixelate	0.979
	t = −0.947
genderMale:corruptionjpeg-compression	1.023
	t = 1.307
Constant	0.839
	t = −24.339***
Observations	2,237,753
Log Likelihood	−953,672.200
Akaike Inf. Crit.	1,907,404.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 26: Gender comparison (odds ratios) with interaction with corruption for the CCD dataset.

	Dependent variable:
	rCE
GenderMale	1.047
	t = 3.909***
GenderOther	0.936
	t = −2.262**
corruptionshot-noise	0.620
	t = −36.971***
corruptionimpulse-noise	0.848
	t = −13.616***
corruptiondefocus-blur	0.103
	t = −96.435***
corruptionglass-blur	0.038
	t = −88.705***
corruptionmotion-blur	0.067
	t = −95.215***
corruptionzoom-blur	4.274
	t = 140.325***
corruptionsnow	0.591
	t = −40.271***
corruptionfrost	0.988
	t = −1.018
corruptionfog	0.085
	t = −96.366***
corruptionbrightness	0.037
	t = −88.401***
corruptioncontrast	0.417
	t = −61.271***
corruptionelastic-transform	0.053
	t = −93.119***
corruptionpixelate	0.148
	t = −93.917***
corruptionjpeg-compression	0.066
	t = −95.103***
GenderMale:corruptionshot-noise	1.111
	t = 5.799***
GenderOther:corruptionshot-noise	1.113
	t = 2.408**
GenderMale:corruptionimpulse-noise	1.001
	t = 0.085
GenderOther:corruptionimpulse-noise	0.993
	t = −0.161
GenderMale:corruptiondefocus-blur	0.905
	t = −2.916***
GenderOther:corruptiondefocus-blur	1.633
	t = 7.009***
GenderMale:corruptionglass-blur	0.965
	t = −0.673
GenderOther:corruptionglass-blur	1.631
	t = 4.560***
GenderMale:corruptionmotion-blur	1.112
	t = 2.694***
GenderOther:corruptionmotion-blur	1.764
	t = 7.025***
GenderMale:corruptionzoom-blur	1.192
	t = 11.867***
GenderOther:corruptionzoom-blur	0.951
	t = −1.390
GenderMale:corruptionsnow	1.034
	t = 1.810*
GenderOther:corruptionsnow	0.974
	t = −0.563
GenderMale:corruptionfrost	0.946
	t = −3.269***
GenderOther:corruptionfrost	0.927
	t = −1.804*
GenderMale:corruptionfog	1.009
	t = 0.258
GenderOther:corruptionfog	1.523
	t = 5.444***
GenderMale:corruptionbrightness	0.910
	t = −1.737*
GenderOther:corruptionbrightness	1.262
	t = 1.931*
GenderMale:corruptioncontrast	1.131
	t = 6.162***
GenderOther:corruptioncontrast	1.153
	t = 2.923***
GenderMale:corruptionelastic-transform	0.952
	t = −1.088
GenderOther:corruptionelastic-transform	1.879
	t = 7.237***
GenderMale:corruptionpixelate	0.938
	t = −2.182**
GenderOther:corruptionpixelate	1.380
	t = 4.993***
GenderMale:corruptionjpeg-compression	0.997
	t = −0.070
GenderOther:corruptionjpeg-compression	1.809
	t = 7.358***
Constant	0.197
	t = −195.801***
Observations	3,216,600
Log Likelihood	−769,351.200
Akaike Inf. Crit.	1,538,792.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 27: Gender comparison (odds ratios) with interaction with corruption for the MIAP dataset.

	Dependent variable:
	γCE
GenderPresentationPredominantly Masculine	1.120 t = 4.908***
GenderPresentationUnknown	1.429 t = 16.648***
corruptionshot-noise	0.979 t = −0.845
corruptionimpulse-noise	0.985 t = −0.592
corruptiondefocus-blur	0.455 t = −26.765***
corruptionglass-blur	0.179 t = −43.807***
corruptionmotion-blur	0.508 t = −23.653***
corruptionzoom-blur	2.422 t = 38.326***
corruptionsnow	0.674 t = −14.622***
corruptionfrost	0.747 t = −11.070***
corruptionfog	0.268 t = −38.449***
corruptionbrightness	0.161 t = −44.725***
corruptioncontrast	0.745 t = −11.138***
corruptionelastic-transform	0.270 t = −38.362***
corruptionpixelate	0.535 t = −22.078***
corruptionjpeg-compression	0.264 t = −38.724***
GenderPresentationPredominantly Masculine:corruptionshot-noise	0.972 t = −0.878
GenderPresentationUnknown:corruptionshot-noise	1.022 t = 0.716
GenderPresentationPredominantly Masculine:corruptionimpulse-noise	1.026 t = 0.798
GenderPresentationUnknown:corruptionimpulse-noise	1.013 t = 0.434
GenderPresentationPredominantly Masculine:corruptiondefocus-blur	0.999 t = −0.033
GenderPresentationUnknown:corruptiondefocus-blur	1.221 t = 5.736***
GenderPresentationPredominantly Masculine:corruptionglass-blur	1.069 t = 1.318
GenderPresentationUnknown:corruptionglass-blur	1.640 t = 11.026***
GenderPresentationPredominantly Masculine:corruptionmotion-blur	1.014 t = 0.377
GenderPresentationUnknown:corruptionmotion-blur	1.151 t = 4.130***
GenderPresentationPredominantly Masculine:corruptionzoom-blur	0.952 t = −1.628
GenderPresentationUnknown:corruptionzoom-blur	0.678 t = −13.759***
GenderPresentationPredominantly Masculine:corruptionsnow	1.130 t = 3.538***
GenderPresentationUnknown:corruptionsnow	1.139 t = 4.049***
GenderPresentationPredominantly Masculine:corruptionfrost	1.093 t = 2.604***
GenderPresentationUnknown:corruptionfrost	1.088 t = 2.660***
GenderPresentationPredominantly Masculine:corruptionfog	1.173 t = 3.675***
GenderPresentationUnknown:corruptionfog	1.471 t = 9.746***
GenderPresentationPredominantly Masculine:corruptionbrightness	1.185 t = 3.293***
GenderPresentationUnknown:corruptionbrightness	1.666 t = 10.979***
GenderPresentationPredominantly Masculine:corruptioncontrast	1.136 t = 3.749***
GenderPresentationUnknown:corruptioncontrast	1.038 t = 1.175
GenderPresentationPredominantly Masculine:corruptionelastic-transform	1.049 t = 1.088
GenderPresentationUnknown:corruptionelastic-transform	1.457 t = 9.503***
GenderPresentationPredominantly Masculine:corruptionpixelate	0.929 t = −1.999**
GenderPresentationUnknown:corruptionpixelate	1.024 t = 0.709
GenderPresentationPredominantly Masculine:corruptionjpeg-compression	1.128 t = 2.740***
GenderPresentationUnknown:corruptionjpeg-compression	1.426 t = 8.893***
Constant	0.254 t = −77.120***
Observations	1,228,938
Log Likelihood	−522,791.900
Akaike Inf. Crit.	1,045,674.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 28: Gender comparison (odds ratios) with interaction with corruption for the UTKFace dataset.

	Dependent variable:
	rCE
genderMale	1.067
	t = 6.513***
corruptionshot-noise	1.205
	t = 18.548***
corruptionimpulse-noise	0.874
	t = −12.869***
corruptiondefocus-blur	0.223
	t = −103.406***
corruptionglass-blur	0.262
	t = −97.098***
corruptionmotion-blur	0.291
	t = −92.484***
corruptionzoom-blur	0.181
	t = −109.707***
corruptionsnow	0.112
	t = −117.447***
corruptionfrost	0.229
	t = −102.441***
corruptionfog	0.082
	t = −118.156***
corruptionbrightness	0.057
	t = −115.677***
corruptioncontrast	0.196
	t = −107.527***
corruptionelastic-transform	0.301
	t = −90.827***
corruptionpixelate	0.129
	t = −116.127***
corruptionjpeg-compression	0.125
	t = −116.448***
genderMale:corruptionshot-noise	0.975
	t = −1.846*
genderMale:corruptionimpulse-noise	0.999
	t = −0.085
genderMale:corruptiondefocus-blur	0.807
	t = −10.430***
genderMale:corruptionglass-blur	0.734
	t = −15.608***
genderMale:corruptionmotion-blur	0.901
	t = −5.627***
genderMale:corruptionzoom-blur	0.804
	t = −9.889***
genderMale:corruptionsnow	1.048
	t = 1.849*
genderMale:corruptionfrost	1.035
	t = 1.768*
genderMale:corruptionfog	0.985
	t = −0.515
genderMale:corruptionbrightness	0.876
	t = −3.807***
genderMale:corruptioncontrast	0.932
	t = −3.369***
genderMale:corruptionelastic-transform	0.916
	t = −4.764***
genderMale:corruptionpixelate	0.873
	t = −5.491***
genderMale:corruptionjpeg-compression	1.005
	t = 0.195
Constant	0.307
	t = −162.844***
Observations	3,314,387
Log Likelihood	−891,903.700
Akaike Inf. Crit.	1,783,867.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 29: Skin Type comparison (odds ratios) for the CCD dataset.

	Dependent variable:
	rCE
FitzDarker	1.162
	t = 38.306***
Constant	0.093
	t = −798.290***
Observations	3,216,600
Log Likelihood	−986,777.800
Akaike Inf. Crit.	1,973,560.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 30: Ethnicity comparison (odds ratios) for the UTKFace dataset with each ethnicity group a reference label.

	Dependent variable:				
			rCE		
	(1)	(2)	(3)	(4)	(5)
ethnicityOther		0.779 t = -31.642^{***}	0.818 t = -23.597^{***}	0.930 t = -8.267^{***}	1.024 t = 2.708^{***}
ethnicityWhite	1.284 t = 31.642^{***}		1.050 t = 9.337^{***}	1.194 t = 30.896^{***}	1.315 t = 48.394^{***}
ethnicityBlack	1.223 t = 23.597^{***}	0.952 t = -9.337^{***}		1.137 t = 19.532^{***}	1.252 t = 34.623^{***}
ethnicityAsian	1.076 t = 8.267^{***}	0.838 t = -30.896^{***}	0.880 t = -19.532^{***}		1.102 t = 14.040^{***}
ethnicityIndian	0.976 t = -2.708^{***}	0.760 t = -48.394^{***}	0.799 t = -34.623^{***}	0.908 t = -14.040^{***}	
Constant	0.085 t = -335.779^{***}	0.109 t = -750.811^{***}	0.104 t = -520.952^{***}	0.092 t = -486.023^{***}	0.083 t = -514.608^{***}
Observations	3,314,987	3,314,987	3,314,987	3,314,987	3,314,987
Log Likelihood	−1,000,414.000	−1,000,414.000	−1,000,414.000	−1,000,414.000	−1,000,414.000
Akaike Inf. Crit.	2,000,839.000	2,000,839.000	2,000,839.000	2,000,839.000	2,000,839.000

Note: *p<0.1; **p<0.05; ***p<0.01

Table 31: Skin Type comparison (odds ratios) with interaction with corruption for the CCD dataset.

	Dependent variable:
	rCE
FitzDarker	1.240
	t = 18.290***
corruptionshot-noise	0.831
	t = −14.098***
corruptionimpulse-noise	0.839
	t = −13.355***
corruptiondefocus-blur	0.112
	t = −87.801***
corruptionglass-blur	0.042
	t = −82.264***
corruptionmotion-blur	0.076
	t = −87.486***
corruptionzoom-blur	4.991
	t = 145.388***
corruptionsnow	0.538
	t = −42.739***
corruptionfrost	0.884
	t = −9.513***
corruptionfog	0.072
	t = −87.186***
corruptionbrightness	0.033
	t = −79.098***
corruptioncontrast	0.410
	t = −57.099***
corruptionelastic-transform	0.055
	t = −85.134***
corruptionpixelate	0.167
	t = −84.350***
corruptionjpeg-compression	0.062
	t = −86.231***
FitzDarker:corruptionshot-noise	0.649
	t = −24.136***
FitzDarker:corruptionimpulse-noise	1.019
	t = 1.078
FitzDarker:corruptiondefocus-blur	0.835
	t = −5.415***
FitzDarker:corruptionglass-blur	0.828
	t = −3.672***
FitzDarker:corruptionmotion-blur	0.927
	t = −1.961**
FitzDarker:corruptionzoom-blur	0.876
	t = −9.033***
FitzDarker:corruptionsnow	1.196
	t = 9.637***
FitzDarker:corruptionfrost	1.149
	t = 8.240***
FitzDarker:corruptionfog	1.361
	t = 8.294***
FitzDarker:corruptionbrightness	1.117
	t = 2.033**
FitzDarker:corruptioncontrast	1.145
	t = 6.748***
FitzDarker:corruptionelastic-transform	0.968
	t = −0.730
FitzDarker:corruptionpixelate	0.795
	t = −8.003***
FitzDarker:corruptionjpeg-compression	1.159
	t = 3.647***
Constant	0.177
	t = −192.404***
Observations	3,216,600
Log Likelihood	−768,244.800
Akaike Inf. Crit.	1,536,550.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 32: Ethnicity comparison (odds ratios) with interaction with corruption for the UTKFace dataset.

	Dependent variable:
	rCE
ethnicityWhite	1.294
	t = 12.668***
ethnicityBlack	1.288
	t = 11.546***
ethnicityAsian	1.014
	t = 0.610
ethnicityIndian	1.163
	t = 6.826***
corruptionshot-noise	1.152
	t = 5.435***
corruptionimpulse-noise	0.858
	t = −5.631***
corruptiondefocus-blur	0.201
	t = −40.499***
corruptionglass-blur	0.249
	t = −37.700***
corruptionmotion-blur	0.274
	t = −36.241***
corruptionzoom-blur	0.181
	t = −41.597***
corruptionsnow	0.123
	t = −44.054***
corruptionfrost	0.226
	t = −39.113***
corruptionfog	0.093
	t = −44.571***
corruptionbrightness	0.064
	t = −45.945***
corruptioncontrast	0.184
	t = −41.420***
corruptionelastic-transform	0.322
	t = −33.318***
corruptionpixelate	0.138
	t = −43.544***
corruptionjpeg-compression	0.133
	t = −43.734***
ethnicityWhite:corruptionshot-noise	1.134
	t = 4.454***
ethnicityBlack:corruptionshot-noise	0.845
	t = −5.489***
ethnicityAsian:corruptionshot-noise	1.098
	t = 2.969***
ethnicityIndian:corruptionshot-noise	0.980
	t = −0.661
ethnicityWhite:corruptionimpulse-noise	1.054
	t = 1.800*
ethnicityBlack:corruptionimpulse-noise	0.952
	t = −1.548
ethnicityAsian:corruptionimpulse-noise	1.072
	t = 2.120**
ethnicityIndian:corruptionimpulse-noise	0.972
	t = −0.901
ethnicityWhite:corruptiondefocus-blur	1.060
	t = 1.370
ethnicityBlack:corruptiondefocus-blur	1.074
	t = 1.563
ethnicityAsian:corruptiondefocus-blur	1.149
	t = 2.936***
ethnicityIndian:corruptiondefocus-blur	0.621
	t = −9.727***
ethnicityWhite:corruptionglass-blur	0.951
	t = −1.270
ethnicityBlack:corruptionglass-blur	0.924
	t = −1.852*
ethnicityAsian:corruptionglass-blur	1.110
	t = 2.363**
ethnicityIndian:corruptionglass-blur	0.556
	t = −12.715***
ethnicityWhite:corruptionmotion-blur	1.036
	t = 0.908
ethnicityBlack:corruptionmotion-blur	1.060
	t = 1.421
ethnicityAsian:corruptionmotion-blur	1.226
	t = 4.789***
ethnicityIndian:corruptionmotion-blur	0.703
	t = −8.122***
ethnicityWhite:corruptionzoom-blur	0.988
	t = −0.261
ethnicityBlack:corruptionzoom-blur	0.931
	t = −1.486
ethnicityAsian:corruptionzoom-blur	0.878
	t = −2.573**
ethnicityIndian:corruptionzoom-blur	0.615
	t = −9.528***
ethnicityWhite:corruptionsnow	0.898
	t = −2.097**
ethnicityBlack:corruptionsnow	1.063
	t = 1.119
ethnicityAsian:corruptionsnow	1.069
	t = 1.163
ethnicityIndian:corruptionsnow	0.744
	t = −5.115***
ethnicityWhite:corruptionfrost	1.011
	t = 0.278
ethnicityBlack:corruptionfrost	1.111
	t = 2.403**
ethnicityAsian:corruptionfrost	1.233
	t = 4.621***
ethnicityIndian:corruptionfrost	0.837
	t = −3.416***
ethnicityWhite:corruptionfog	0.844
	t = −2.920***
ethnicityBlack:corruptionfog	0.976
	t = −0.389
ethnicityAsian:corruptionfog	1.010
	t = 0.147
ethnicityIndian:corruptionfog	0.692
	t = −5.606***
ethnicityWhite:corruptionbrightness	0.828
	t = −2.771***
ethnicityBlack:corruptionbrightness	0.918
	t = −1.170
ethnicityAsian:corruptionbrightness	0.906
	t = −1.276
ethnicityIndian:corruptionbrightness	0.600
	t = −6.484***
ethnicityWhite:corruptioncontrast	1.000
	t = 0.002
ethnicityBlack:corruptioncontrast	1.007
	t = 0.142
ethnicityAsian:corruptioncontrast	1.272
	t = 4.973***
ethnicityIndian:corruptioncontrast	0.918
	t = −1.763*
ethnicityWhite:corruptionelastic-transform	0.882
	t = −3.391***
ethnicityBlack:corruptionelastic-transform	0.905
	t = −2.521**
ethnicityAsian:corruptionelastic-transform	1.007
	t = 0.161
ethnicityIndian:corruptionelastic-transform	0.770
	t = −6.400***
ethnicityWhite:corruptionpixelate	0.916
	t = −1.782*
ethnicityBlack:corruptionpixelate	0.970
	t = −0.571
ethnicityAsian:corruptionpixelate	0.878
	t = −2.317**
ethnicityIndian:corruptionpixelate	0.580
	t = −9.553***
ethnicityWhite:corruptionjpeg-compression	1.051
	t = 1.011
ethnicityBlack:corruptionjpeg-compression	0.982
	t = −0.349
ethnicityAsian:corruptionjpeg-compression	0.863
	t = −2.572**
ethnicityIndian:corruptionjpeg-compression	0.687
	t = −6.634***
Constant	0.265
	t = −70.758***
Observations	3,314,987
Log-Likelihood	−889,589.400
Akaike Inf. Crit.	1,779,329.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 33: Lighting comparison (odds ratios) for the CCD dataset.

	Dependent variable:
	rCE
lightingDark	1.676
	t = 130.051***
Constant	0.085
	t = −997.540***
Observations	3,216,600
Log Likelihood	−979,343.200
Akaike Inf. Crit.	1,958,690.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 34: Lighting comparison (odds ratios) with interaction with corruption for the CCD dataset.

	Dependent variable:
	rCE
lightingDark	2.413 t = 74.079***
corruptionshot-noise	0.895 t = −9.974***
corruptionimpulse-noise	0.816 t = −17.890***
corruptiondefocus-blur	0.148 t = −98.245***
corruptionglass-blur	0.058 t = −97.544***
corruptionmotion-blur	0.108 t = −100.424***
corruptionzoom-blur	6.450 t = 201.159***
corruptionsnow	0.391 t = −68.383***
corruptionfrost	0.718 t = −28.401***
corruptionfog	0.073 t = −99.522***
corruptionbrightness	0.052 t = −96.273***
corruptioncontrast	0.510 t = −52.973***
corruptionelastic-transform	0.081 t = −100.105***
corruptionpixelate	0.229 t = −89.619***
corruptionjpeg-compression	0.085 t = −100.277***
lightingDark:corruptionshot-noise	0.420 t = −45.819***
lightingDark:corruptionimpulse-noise	1.082 t = 4.577***
lightingDark:corruptiondefocus-blur	0.322 t = −29.974***
lightingDark:corruptionglass-blur	0.272 t = −21.101***
lightingDark:corruptionmotion-blur	0.314 t = −26.376***
lightingDark:corruptionzoom-blur	0.390 t = −61.956***
lightingDark:corruptionsnow	2.220 t = 42.119***
lightingDark:corruptionfrost	1.856 t = 36.243***
lightingDark:corruptionfog	1.331 t = 8.049***
lightingDark:corruptionbrightness	0.338 t = −18.013***
lightingDark:corruptioncontrast	0.699 t = −17.868***
lightingDark:corruptionelastic-transform	0.287 t = −24.232***
lightingDark:corruptionpixelate	0.239 t = −41.315***
lightingDark:corruptionjpeg-compression	0.558 t = −14.311***
Constant	0.146 t = −249.377***
Observations	3,216,600
Log Likelihood	−748,579.900
Akaike Inf. Crit.	1,497,220.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 35: Interaction (odds ratio) of Lighting and Skin Type for the CCD dataset.

	Dependent variable:		
	rCE		
	Interaction	Just dark Lighting	Just light Lighting
	(1)	(2)	(3)
lightingDark	1.716		
	t = 82.310***		
FitzDarker	1.090	1.030	1.090
	t = 17.447***	t = 4.369***	t = 17.447***
lightingDark:FitzDarker	0.945		
	t = −6.868***		
Constant	0.081	0.140	0.081
	t = −708.480***	t = −356.507***	t = −708.480***
Observations	3,216,600	948,600	2,268,000
Log Likelihood	−979,181.300	−356,677.900	−622,503.400
Akaike Inf. Crit.	1,958,371.000	713,359.700	1,245,011.000

Note: *p<0.1; **p<0.05; ***p<0.01

Table 36: Interaction (odds ratio) of Lighting and Age for the CCD dataset.

	Dependent variable:		
	rCE		
	Interaction	Just dark Lighting	Just light Lighting
	(1)	(2)	(3)
lightingDark	1.792 t = 105.735***		
Age45-64	1.183 t = 30.918***	1.127 t = 17.049***	1.183 t = 30.918***
Age65+	1.312 t = 38.111***	1.241 t = 17.731***	1.312 t = 38.111***
lightingDark:Age45-64	0.952 t = −5.522***		
lightingDark:Age65+	0.946 t = −3.939***		
Constant	0.076 t = −657.440***	0.136 t = −514.889***	0.076 t = −657.440***
Observations	3,216,600	948,600	2,268,000
Log Likelihood	−978,216.300	−356,434.600	−621,781.600
Akaike Inf. Crit.	1,956,445.000	712,875.200	1,243,569.000
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 37: Interaction (odds ratio) of Lighting and Age (as a numeric variable) for the CCD dataset.

	Dependent variable:
	rCE
lightingDark	1.782
	t = 53.960***
Age_Numeric	1.005
	t = 40.613***
lightingDark:Age_Numeric	0.999
	t = −3.014***
Constant	0.067
	t = −411.641***
Observations	3,216,600
Log Likelihood	−978,233.700
Akaike Inf. Crit.	1,956,475.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 38: Interaction (odds ratio) of Lighting and Gender for the CCD dataset.

	Dependent variable:		
	rCE		
	Interaction	Just dark Lighting	Just light Lighting
	(1)	(2)	(3)
lightingDark	1.741 t = 97.740***		
GenderMale	1.138 t = 25.421***	1.060 t = 9.219***	1.138 t = 25.421***
GenderOther	1.104 t = 9.008***	1.173 t = 7.409***	1.104 t = 9.008***
lightingDark:GenderMale	0.931 t = −8.827***		
lightingDark:GenderOther	1.062 t = 2.504**		
Constant	0.079 t = −687.244***	0.138 t = −458.693***	0.079 t = −687.244***
Observations	3,216,600	948,600	2,268,000
Log Likelihood	−978,953.700	−356,626.200	−622,327.500
Akaike Inf. Crit.	1,957,919.000	713,258.500	1,244,661.000
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 39: Interaction (odds ratio) of Age and Gender for the CCD dataset.

	Dependent variable:			
	rCE			
	Interaction	Just 19-45	Just 45-64	Just 65+
	(1)	(2)	(3)	(4)
Age45-64	1.010 t = 1.642			
Age65+	1.057 t = 6.501***			
GenderMale	1.048 t = 8.270***	1.048 t = 8.270***	1.124 t = 18.210***	1.135 t = 11.891***
GenderOther	0.978 t = −1.877*	0.978 t = −1.877*	0.941 t = −3.201***	3.430 t = 25.298***
Age45-64:GenderMale	1.072 t = 8.145***			
Age65+:GenderMale	1.083 t = 6.593***			
Age45-64:GenderOther	0.962 t = −1.736*			
Age65+:GenderOther	3.506 t = 25.033***			
Constant	0.096 t = −581.789***	0.096 t = −581.789***	0.097 t = −512.021***	0.102 t = −306.536***
Observations	3,216,600	1,623,450	1,188,750	404,400
Log Likelihood	−986,810.400	−489,181.800	−367,860.900	−129,767.700
Akaike Inf. Crit.	1,973,639.000	978,369.600	735,727.900	259,541.400

Note: *p<0.1; **p<0.05; ***p<0.01

Table 40: Interaction (odds ratio) of Age, Skin Type, and Gender for the CCD dataset.

	Dependent variable:
	rCE
FitzDarker	1.230 t = 24.890***
Age45-64	1.133 t = 13.192***
Age65+	1.155 t = 11.878***
GenderMale	0.965 t = −3.694***
GenderOther	0.993 t = −0.364
FitzDarker:Age45-64	0.829 t = −15.192***
FitzDarker:Age65+	0.879 t = −7.557***
FitzDarker:GenderMale	1.128 t = 10.118***
FitzDarker:GenderOther	0.969 t = −1.287
Age45-64:GenderMale	1.103 t = 7.259***
Age65+:GenderMale	1.198 t = 10.728***
Age45-64:GenderOther	0.782 t = −5.346***
Age65+:GenderOther	3.418 t = 23.871***
FitzDarker:Age45-64:GenderMale	0.984 t = −0.903
FitzDarker:Age65+:GenderMale	0.860 t = −6.079***
FitzDarker:Age45-64:GenderOther	1.314 t = 5.154***
FitzDarker:Age65+:GenderOther	
Constant	0.085 t = −375.465***
Observations	3,216,600
Log Likelihood	−985,574.900
Akaike Inf. Crit.	1,971,184.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 41: Interaction (odds ratio) of Age (as a numeric variable) and Gender for the CCD dataset.

	Dependent variable:
	rCE
Age_Numeric	1.001
	t = 8.277***
GenderMale	0.964
	t = −3.213***
GenderOther	0.966
	t = −2.628***
Age_Numeric:GenderMale	1.003
	t = 11.259***
Age_Numeric:GenderOther	1.008
	t = 15.957***
Constant	0.092
	t = −298.173***
Observations	3,216,600
Log Likelihood	−986,768.500
Akaike Inf. Crit.	1,973,549.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 42: Interaction (odds ratio) of Age and Skin Type for the CCD dataset.

	Dependent variable:			
	rCE			
	Interaction	Just 19-45	Just 45-64	Just 65+
	(1)	(2)	(3)	(4)
Age45-64	1.183 t = 25.478***			
Age65+	1.268 t = 28.473***			
FitzDarker	1.298 t = 45.414***	1.298 t = 45.414***	1.071 t = 10.740***	1.075 t = 6.763***
Age45-64:FitzDarker	0.825 t = −22.526***			
Age65+:FitzDarker	0.828 t = −15.540***			
Constant	0.084 t = −537.849***	0.084 t = −537.849***	0.099 t = −491.532***	0.106 t = −323.904***
Observations	3,216,600	1,623,450	1,188,750	404,400
Log Likelihood	−986,215.000	−488,172.200	−367,989.600	−130,053.200
Akaike Inf. Crit.	1,972,442.000	976,348.400	735,983.100	260,110.500
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 43: Interaction (odds ratio) of Age (as a numeric variable) and Skin Type for the CCD dataset.

	Dependent variable:
	rCE
Age_Numeric	1.005
	t = 31.573***
FitzDarker	1.357
	t = 29.671***
Age_Numeric:FitzDarker	0.997
	t = −15.120***
Constant	0.074
	t = −326.571***
Observations	3,216,600
Log Likelihood	−986,196.600
Akaike Inf. Crit.	1,972,401.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 44: Interaction (odds ratio) of Gender and Skin Type for the CCD dataset.

	Dependent variable:			
	rCE			
	Interaction	Just Male	Just Female	Just Other
	(1)	(2)	(3)	(4)
GenderMale	1.044 t = 7.146***			
GenderOther	0.903 t = −5.850***			
FitzDarker	1.119 t = 19.829***	1.204 t = 32.947***	1.119 t = 19.829***	1.262 t = 11.526***
GenderMale:FitzDarker	1.076 t = 9.224***			
GenderOther:FitzDarker	1.128 t = 5.757***			
Constant	0.091 t = −558.827***	0.095 t = −550.458***	0.091 t = −558.827***	0.083 t = −147.736***
Observations	3,216,600	1,488,150	1,583,550	144,900
Log Likelihood	−986,468.800	−469,035.600	−474,279.700	−43,153.540
Akaike Inf. Crit.	1,972,950.000	938,075.200	948,563.300	86,311.070

Note: *p<0.1; **p<0.05; ***p<0.01

Table 45: Interaction (odds ratio) of Age and Gender for the Adience dataset.

	Dependent variable:								
					rCE				
	Interaction	Just 0-2	Just 3-7	Just 8-14	Just 15-24	Just 25-35	Just 36-45	Just 46-59	Just 60+
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
age_group3-7	0.872 t = −13.412***								
age_group8-14	0.976 t = −2.460**								
age_group15-24	0.952 t = −4.731***								
age_group25-35	1.219 t = 20.707***								
age_group36-45	1.128 t = 11.808***								
age_group46-59	1.085 t = 6.412***								
age_group60+	1.132 t = 9.760***								
genderMale	0.941 t = −5.337***	0.941 t = −5.337***	1.028 t = 2.865***	0.871 t = −14.881***	1.156 t = 14.435***	1.066 t = 8.822***	1.067 t = 7.959***	1.084 t = 5.958***	1.213 t = 14.570***
age_group3-7:genderMale	1.092 t = 5.930***								
age_group8-14:genderMale	0.925 t = −5.280***								
age_group15-24:genderMale	1.228 t = 13.542***								
age_group25-35:genderMale	1.132 t = 9.223***								
age_group36-45:genderMale	1.134 t = 8.973***								
age_group46-59:genderMale	1.152 t = 7.995***								
age_group60+:genderMale	1.289 t = 14.528***								
Constant	0.224 t = −185.350***	0.224 t = −185.350***	0.195 t = −259.643***	0.218 t = −262.880***	0.213 t = −233.163***	0.273 t = −252.805***	0.252 t = −222.062***	0.243 t = −143.491***	0.253 t = −139.692***
Observations	2,237,753	210,000	323,543	342,858	268,347	449,989	373,483	135,288	134,245
Log Likelihood	−1,094,788.000	−98,445.690	−144,836.500	−156,906.000	−128,298.100	−236,970.700	−190,725.500	−68,082.690	−70,523.140
Akaike Inf. Crit.	2,189,608.000	196,895.400	289,677.000	313,816.000	256,600.200	473,945.400	381,454.900	136,169.400	141,050.300
Note:									*p<0.1; **p<0.05; ***p<0.01

Table 46: Interaction (odds ratio) of Age and Gender for the MIAP dataset.

	Dependent variable:						
				rCE			
	Interaction	Just Young	Just Middle	Just Old	Just Feminine	Just Masculine	Just Unknown
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
AgePresentationMiddle	0.675 t = −19.644***				0.675 t = −19.644***	0.809 t = −19.982***	0.957 t = −0.932
AgePresentationOlder	1.270 t = 15.832***				1.270 t = 15.832***	1.280 t = 20.790***	1.852 t = 58.289***
AgePresentationUnknown	0.856 t = −13.842***						0.856 t = −13.842***
GenderPresentationPredominantly Masculine	1.163 t = 17.934***	1.163 t = 17.934***	1.395 t = 15.833***	1.172 t = 9.211***			
GenderPresentationUnknown	1.190 t = 15.478***	1.190 t = 15.478***	1.688 t = 10.466***	1.735 t = 37.711***			
AgePresentationMiddle:GenderPresentationPredominantly Masculine	1.199 t = 8.019***						
AgePresentationOlder:GenderPresentationPredominantly Masculine	1.008 t = 0.394						
AgePresentationUnknown:GenderPresentationPredominantly Masculine							
AgePresentationMiddle:GenderPresentationUnknown	1.418 t = 6.820***						
AgePresentationOlder:GenderPresentationUnknown	1.458 t = 20.468***						
AgePresentationUnknown:GenderPresentationUnknown							
Constant	0.159 t = −299.959***	0.159 t = −299.959***	0.108 t = −116.917***	0.202 t = −115.744***	0.159 t = −299.959***	0.185 t = −290.356***	0.190 t = −176.549***
Observations	1,228,938	534,150	147,075	322,713	293,850	397,415	537,673
Log Likelihood	−545,873.700	−224,575.000	−54,955.920	−175,345.900	−117,066.700	−170,469.000	−258,338.000
Akaike Inf. Crit.	1,091,767.000	449,155.900	109,917.800	350,697.900	234,139.400	340,944.100	516,684.000

Note: *p<0.1; **p<0.05; ***p<0.01

Table 47: Interaction (odds ratio) of Age and Gender for the UTKFace dataset.

	Dependent variable:				
			rCE		
	Interaction	Just 0-18	Just 19-45	Just 45-64	Just 65+
	(1)	(2)	(3)	(4)	(5)
Age19-45	0.855 t = -22.893^{***}				
Age45-64	0.964 t = -3.723^{***}				
Age65+	1.287 t = 25.011^{***}				
genderMale	0.945 t = -6.807^{***}	0.945 t = -6.807^{***}	1.009 t = 1.710^{*}	1.036 t = 3.620^{***}	0.914 t = -7.823^{***}
Age19-45:genderMale	1.069 t = 6.625^{***}				
Age45-64:genderMale	1.096 t = 7.198^{***}				
Age65+:genderMale	0.968 t = -2.298^{**}				
Constant	0.105 t = -395.638^{***}	0.105 t = -395.638^{***}	0.090 t = -631.394^{***}	0.101 t = -282.362^{***}	0.135 t = -239.992^{***}
Observations	3,314,387	678,662	1,728,150	606,600	300,975
Log Likelihood	−1,000,139.000	−209,933.600	−493,906.800	−189,325.500	−106,973.500
Akaike Inf. Crit.	2,000,295.000	419,871.200	987,817.600	378,654.900	213,951.000

Note: *p<0.1; **p<0.05; ***p<0.01

Table 48: Interaction (odds ratio) of Age and Ethnicity for the UTKFace dataset.

	Dependent variable:				
			rCE		
	Interaction	Just 0-18	Just 19-45	Just 45-64	Just 65+
	(1)	(2)	(3)	(4)	(5)
Age19-45	0.780 t = −15.801***				
Age45-64	0.781 t = −7.867***				
Age65+	1.800 t = 8.801***				
ethnicityWhite	1.215 t = 14.166***	1.215 t = 14.166***	1.191 t = 15.760***	1.412 t = 11.706***	0.719 t = −5.010***
ethnicityBlack	0.882 t = −5.821***	0.882 t = −5.821***	1.328 t = 25.779***	1.357 t = 9.849***	0.731 t = −4.640***
ethnicityAsian	1.004 t = 0.283	1.004 t = 0.283	1.084 t = 6.786***	1.080 t = 2.177**	0.683 t = −5.619***
ethnicityIndian	0.607 t = −26.663***	0.607 t = −26.663***	1.062 t = 5.245***	1.221 t = 6.512***	0.741 t = −4.335***
Age19-45:ethnicityWhite	0.980 t = −1.119				
Age45-64:ethnicityWhite	1.162 t = 4.618***				
Age65+:ethnicityWhite	0.591 t = −7.795***				
Age19-45:ethnicityBlack	1.505 t = 16.911***				
Age45-64:ethnicityBlack	1.539 t = 11.411***				
Age65+:ethnicityBlack	0.828 t = −2.656***				
Age19-45:ethnicityAsian	1.079 t = 3.996***				
Age45-64:ethnicityAsian	1.076 t = 1.896*				
Age65+:ethnicityAsian	0.680 t = −5.548***				
Age19-45:ethnicityIndian	1.751 t = 25.455***				
Age45-64:ethnicityIndian	2.012 t = 19.454***				
Age65+:ethnicityIndian	1.221 t = 2.785***				
Constant	0.100 t = −186.123***	0.100 t = −186.123***	0.078 t = −262.518***	0.078 t = −88.297***	0.180 t = −26.173***
Observations	3,314,537	678,662	1,728,150	606,750	300,975
Log Likelihood	−998,180.900	−208,711.700	−493,285.800	−189,196.900	−106,986.500
Akaike Inf. Crit.	1,996,402.000	417,433.400	986,581.500	378,403.700	213,983.100

Note: *p<0.1; **p<0.05; ***p<0.01

Table 49: Interaction (odds ratio) of Gender and Ethnicity for the UTKFace dataset.

	Dependent variable:		
	rCE		
	Interaction	Just Male	Just Female
	(1)	(2)	(3)
genderMale	0.822 t = -13.125^{***}		
ethnicityWhite	1.198 t = 17.206^{***}	1.420 t = 28.782^{***}	1.198 t = 17.206^{***}
ethnicityBlack	1.074 t = 6.190^{***}	1.424 t = 27.375^{***}	1.074 t = 6.190^{***}
ethnicityAsian	0.957 t = -3.732^{***}	1.251 t = 16.519^{***}	0.957 t = -3.732^{***}
ethnicityIndian	0.908 t = -8.045^{***}	1.084 t = 6.121^{***}	0.908 t = -8.045^{***}
genderMale:ethnicityWhite	1.186 t = 10.601^{***}		
genderMale:ethnicityBlack	1.326 t = 16.291^{***}		
genderMale:ethnicityAsian	1.306 t = 14.934^{***}		
genderMale:ethnicityIndian	1.194 t = 9.945^{***}		
Constant	0.093 t = -249.112^{***}	0.076 t = -224.112^{***}	0.093 t = -249.112^{***}
Observations	3,314,387	1,729,045	1,585,342
Log Likelihood	−1,000,032.000	−522,244.200	−477,787.900
Akaike Inf. Crit.	2,000,084.000	1,044,498.000	955,585.700
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 50: Corruption comparison (odds ratios) of Gender Prediction Error by dataset.

	Dependent variable:			
	pred_gender_iswrong			
	Adience	CCD	MIAP	UTKFace
	(1)	(2)	(3)	(4)
corruptionclean	0.006	0.051	0.314	0.079
	t = −48.610***	t = −91.515***	t = −33.546***	t = −98.300***
corruptiongaussian-noise	0.833	0.402	0.757	0.433
	t = −24.808***	t = −132.005***	t = −20.900***	t = −127.837***
corruptionshot-noise	0.842	0.345	0.771	0.492
	t = −23.327***	t = −148.778***	t = −19.582***	t = −110.901***
corruptionimpulse-noise	0.841	0.400	0.787	0.409
	t = −23.532***	t = −132.316***	t = −18.052***	t = −134.944***
corruptiondefocus-blur	0.027	0.066	0.527	0.221
	t = −158.028***	t = −209.602***	t = −46.258***	t = −193.277***
corruptionglass-blur	0.040	0.052	0.371	0.221
	t = −169.313***	t = −205.248***	t = −66.909***	t = −193.074***
corruptionmotion-blur	0.038	0.063	0.558	0.200
	t = −167.885***	t = −208.898***	t = −42.520***	t = −199.330***
corruptionzoom-blur	0.017	1.468	1.710	0.152
	t = −142.312***	t = 60.319***	t = 39.282***	t = −211.972***
corruptionsnow	0.414	0.380	0.696	0.148
	t = −109.677***	t = −138.332***	t = −27.012***	t = −212.783***
corruptionfrost	0.645	0.374	0.721	0.214
	t = −58.425***	t = −139.927***	t = −24.495***	t = −195.199***
corruptionfog	0.319	0.095	0.461	0.134
	t = −133.671***	t = −212.022***	t = −54.576***	t = −215.671***
corruptionbrightness	0.037	0.060	0.382	0.092
	t = −166.731***	t = −208.048***	t = −65.286***	t = −220.294***
corruptioncontrast	0.311	0.230	0.768	0.206
	t = −135.567***	t = −183.393***	t = −19.856***	t = −197.658***
corruptionelastic-transform	0.106	0.057	0.404	0.186
	t = −180.467***	t = −207.143***	t = −62.337***	t = −203.417***
corruptionpixelate	0.009	0.092	0.577	0.112
	t = −122.500***	t = −212.058***	t = −40.195***	t = −218.991***
corruptionjpeg-compression	0.277	0.073	0.422	0.125
	t = −144.384***	t = −210.944***	t = −59.772***	t = −217.285***
Observations	1,133,844	1,556,328	350,284	1,679,524
Log Likelihood	−449,312.600	−619,803.800	−226,368.400	−746,679.000
Akaike Inf. Crit.	898,657.300	1,239,640.000	452,768.800	1,493,390.000

Note: *p<0.1; **p<0.05; ***p<0.01

Table 51: Gender prediction by Age comparison (odds ratios) for the Adience dataset with each age group a reference label.

	Dependent variable:							
	pred_gender_iswrong							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
age_group0-2		0.979 t = −2.261**	1.084 t = 8.739***	1.433 t = 35.729***	1.719 t = 58.716***	1.538 t = 45.633***	1.445 t = 30.001***	1.156 t = 12.299***
age_group3-7	1.021 t = 2.261**		1.107 t = 12.444***	1.463 t = 41.800***	1.755 t = 68.918***	1.571 t = 53.742***	1.476 t = 33.838***	1.180 t = 15.105***
age_group8-14	0.923 t = −8.739***	0.904 t = −12.444***		1.322 t = 30.712***	1.586 t = 56.584***	1.419 t = 41.740***	1.334 t = 25.053***	1.067 t = 5.889***
age_group15-24	0.698 t = −35.729***	0.683 t = −41.800***	0.756 t = −30.712***		1.199 t = 19.945***	1.073 t = 7.594***	1.009 t = 0.711	0.807 t = −18.339***
age_group25-35	0.582 t = −58.716***	0.570 t = −68.918***	0.631 t = −56.584***	0.834 t = −19.945***		0.895 t = −13.190***	0.841 t = −15.041***	0.673 t = −36.082***
age_group36-45	0.650 t = −45.633***	0.637 t = −53.742***	0.705 t = −41.740***	0.932 t = −7.594***	1.117 t = 13.190***		0.940 t = −5.322***	0.752 t = −25.568***
age_group46-59	0.692 t = −30.001***	0.678 t = −33.838***	0.750 t = −25.053***	0.991 t = −0.711	1.189 t = 15.041***	1.064 t = 5.322***		0.800 t = −16.357***
age_group60+	0.865 t = −12.299***	0.847 t = −15.105***	0.937 t = −5.889***	1.240 t = 18.339***	1.487 t = 36.082***	1.331 t = 25.568***	1.250 t = 16.357***	
Constant	0.322 t = −157.523***	0.329 t = −192.949***	0.297 t = −211.063***	0.225 t = −211.600***	0.188 t = −289.533***	0.210 t = −255.603***	0.223 t = −150.709***	0.279 t = −136.638***
Observations	1,118,925	1,118,925	1,118,925	1,118,925	1,118,925	1,118,925	1,118,925	1,118,925
Log Likelihood	−556,687.100	−556,687.100	−556,687.100	−556,687.100	−556,687.100	−556,687.100	−556,687.100	−556,687.100
Akaike Inf. Crit.	1,113,390.000	1,113,390.000	1,113,390.000	1,113,390.000	1,113,390.000	1,113,390.000	1,113,390.000	1,113,390.000

Note: *p<0.1; **p<0.05; ***p<0.01

Table 52: Gender prediction by Age comparison (odds ratios) for the CCD dataset with each age group a reference label.

	Dependent variable:		
	pred_gender_iswrong		
	(1)	(2)	(3)
Age19-45		1.093 t = 19.369***	1.035 t = 5.208***
Age45-64	0.915 t = −19.369***		0.947 t = −8.075***
Age65+	0.967 t = −5.208***	1.056 t = 8.075***	
Constant	0.226 t = −502.801***	0.207 t = −450.595***	0.218 t = −261.764***
Observations	1,535,850	1,535,850	1,535,850
Log Likelihood	−720,624.000	−720,624.000	−720,624.000
Akaike Inf. Crit.	1,441,254.000	1,441,254.000	1,441,254.000
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 53: Gender prediction by Age (as a numeric variable) comparison (odds ratios) for the CCD dataset with each age group a reference label.

	Dependent variable:
	pred_gender_iswrong
Age_Numeric	0.998
	t = -15.711^{***}
Constant	0.238
	t = -241.062^{***}
Observations	1,535,850
Log Likelihood	−720,688.900
Akaike Inf. Crit.	1,441,382.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 54: Gender prediction by Age comparison (odds ratios) for the MIAP dataset with each age group a reference label.

	Dependent variable:		
	pred_gender_iswrong		
	(1)	(2)	(3)
AgePresentationYoung		1.370 t = 32.825***	0.127 t = −171.092***
AgePresentationMiddle	0.730 t = −32.825***		0.093 t = −169.365***
AgePresentationOlder	7.852 t = 171.092***	10.756 t = 169.365***	
Constant	0.490 t = −158.830***	0.358 t = −121.263***	3.850 t = 120.606***
Observations	345,675	345,675	345,675
Log Likelihood	−208,814.200	−208,814.200	−208,814.200
Akaike Inf. Crit.	417,634.300	417,634.300	417,634.300
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 55: Gender prediction by Age comparison (odds ratios) for the UTKFace dataset with each age group a reference label.

	Dependent variable:			
	pred_gender_iswrong			
	(1)	(2)	(3)	(4)
Age0-18		4.672	4.030	2.739
		t = 313.886***	t = 215.173***	t = 132.490***
Age19-45	0.214		0.863	0.586
	t = −313.886***		t = −23.097***	t = −70.837***
Age45-64	0.248	1.159		0.680
	t = −215.173***	t = 23.097***		t = −44.695***
Age65+	0.365	1.705	1.471	
	t = −132.490***	t = 70.837***	t = 44.695***	
Constant	0.596	0.128	0.148	0.218
	t = −145.801***	t = −606.284***	t = −352.628***	t = −226.675***
Observations	1,657,425	1,657,425	1,657,425	1,657,425
Log Likelihood	−716,651.200	−716,651.200	−716,651.200	−716,651.200
Akaike Inf. Crit.	1,433,310.000	1,433,310.000	1,433,310.000	1,433,310.000
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 56: Gender prediction by Age (as a numeric variable) comparison (odds ratios) for the UTKFace dataset with each age group a reference label.

	Dependent variable:
	pred_gender_iswrong
Age_Numeric	0.979
	t = -190.832^{***}
Constant	0.402
	t = -247.245^{***}
Observations	1,657,425
Log Likelihood	$-749{,}465.600$
Akaike Inf. Crit.	1,498,935.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 57: Gender prediction by Gender comparison (odds ratios) for the Adience dataset with each gender group a reference label.

	Dependent variable:	
	pred_gender_iswrong	
	(1)	(2)
genderMale	1.730	
	t = 114.552***	
genderFemale		0.578
		t = −114.552***
Constant	0.189	0.326
	t = −464.936***	t = −353.504***
Observations	1,118,925	1,118,925
Log Likelihood	−554,151.600	−554,151.600
Akaike Inf. Crit.	1,108,307.000	1,108,307.000
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 58: Gender prediction by Gender comparison (odds ratios) for the CCD dataset with each gender group a reference label.

	Dependent variable:	
	pred_gender_iswrong	
	(1)	(2)
GenderMale	1.090	
	t = 20.434***	
GenderFemale		0.918
		t = −20.434***
Constant	0.208	0.227
	t = −527.125***	t = −496.482***
Observations	1,535,850	1,535,850
Log Likelihood	−720,603.600	−720,603.600
Akaike Inf. Crit.	1,441,211.000	1,441,211.000
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 59: Gender prediction by Gender comparison (odds ratios) for the MIAP dataset with each gender group a reference label.

	Dependent variable:	
	pred_gender_iswrong	
	(1)	(2)
GenderPresentationPredominantly Masculine	1.096	
	t = 12.860***	
GenderPresentationPredominantly Feminine		0.913
		t = −12.860***
Constant	0.584	0.640
	t = −99.360***	t = −96.988***
Observations	345,675	345,675
Log Likelihood	−229,667.000	−229,667.000
Akaike Inf. Crit.	459,338.000	459,338.000
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 60: Gender prediction by Gender comparison (odds ratios) for the UTKFace dataset with each gender group a reference label.

	Dependent variable:	
	pred_gender_iswrong	
	(1)	(2)
genderMale	2.291	
	t = 190.786***	
genderFemale		0.437
		t = −190.786***
Constant	0.131	0.299
	t = −579.393***	t = −472.514***
Observations	1,657,425	1,657,425
Log Likelihood	−750,147.900	−750,147.900
Akaike Inf. Crit.	1,500,300.000	1,500,300.000
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 61: Gender prediction by Skin Type comparison (odds ratios) for the CCD dataset.

	Dependent variable:
	pred_gender_iswrong
FitzDarker	1.107
	t = 24.021***
Constant	0.206
	t = −497.992***
Observations	1,535,850
Log Likelihood	−720,523.100
Akaike Inf. Crit.	1,441,050.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 62: Gender prediction by Ethnicity comparison (odds ratios) for the UTKFace dataset with each ethnicity group a reference label.

	Dependent variable:				
	pred_gender_iswrong				
	(1)	(2)	(3)	(4)	(5)
ethnicityOther		0.968 t = −4.028***	1.124 t = 13.192***	0.715 t = −38.539***	1.204 t = 20.591***
ethnicityWhite	1.033 t = 4.028***		1.162 t = 25.323***	0.738 t = −53.719***	1.243 t = 35.849***
ethnicityBlack	0.889 t = −13.192***	0.861 t = −25.323***		0.636 t = −67.028***	1.070 t = 9.542***
ethnicityAsian	1.399 t = 38.539***	1.354 t = 53.719***	1.573 t = 67.028***		1.684 t = 75.527***
ethnicityIndian	0.831 t = −20.591***	0.804 t = −35.849***	0.934 t = −9.542***	0.594 t = −75.527***	
Constant	0.208 t = −212.207***	0.215 t = −471.968***	0.185 t = −341.174***	0.291 t = −267.290***	0.173 t = −341.865***
Observations	1,657,425	1,657,425	1,657,425	1,657,425	1,657,425
Log Likelihood	−765,963.100	−765,963.100	−765,963.100	−765,963.100	−765,963.100
Akaike Inf. Crit.	1,531,936.000	1,531,936.000	1,531,936.000	1,531,936.000	1,531,936.000

Note: *p<0.1; **p<0.05; ***p<0.01

Table 63: Gender Prediction by Lighting comparison (odds ratios) for the CCD dataset.

	Dependent variable:
	pred_gender_iswrong
lightingDark	1.302
	t = 59.258***
Constant	0.200
	t = −620.858***
Observations	1,535,850
Log Likelihood	−719,086.800
Akaike Inf. Crit.	1,438,178.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 64: Corruption comparison of Age Estimation Error by dataset.

	Dependent variable:	
	diff_pred_age	
	CCD	UTKFace
	(1)	(2)
corruptiongaussian-noise	6.653***	1.101***
	(0.105)	(0.098)
corruptionshot-noise	5.075***	0.917***
	(0.105)	(0.099)
corruptionimpulse-noise	6.786***	1.644***
	(0.106)	(0.098)
corruptiondefocus-blur	0.544***	0.775***
	(0.105)	(0.097)
corruptionglass-blur	0.364***	−1.444***
	(0.105)	(0.097)
corruptionmotion-blur	1.583***	2.559***
	(0.105)	(0.097)
corruptionzoom-blur	4.402***	0.902***
	(0.113)	(0.097)
corruptionsnow	2.873***	−1.034***
	(0.106)	(0.097)
corruptionfrost	0.955***	1.060***
	(0.107)	(0.097)
corruptionfog	3.332***	1.164***
	(0.105)	(0.097)
corruptionbrightness	0.843***	1.022***
	(0.105)	(0.097)
corruptioncontrast	4.769***	2.971***
	(0.106)	(0.097)
corruptionelastic-transform	−0.473***	−0.670***
	(0.105)	(0.097)
corruptionpixelate	0.600***	0.720***
	(0.106)	(0.097)
corruptionjpeg-compression	−0.930***	−0.657***
	(0.105)	(0.097)
Constant	7.580***	2.996***
	(0.096)	(0.088)
Observations	1,474,920	1,590,651
Log Likelihood	−5,961,911.000	−6,355,681.000
Akaike Inf. Crit.	11,923,855.000	12,711,393.000
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 65: Age Estimation by Age (as a numeric variable) comparison for the CCD dataset with each age group a reference label.

	Dependent variable:
	diff_pred_age
Age_Numeric	0.594^{***}
	(0.001)
Constant	-16.109^{***}
	(0.025)
Observations	1,454,363
Log Likelihood	−5,459,682.000
Akaike Inf. Crit.	10,919,368.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 66: Age Estimation by Age (as a numeric variable) comparison for the UTKFace dataset with each age group a reference label.

	Dependent variable:
	diff_pred_age
Age_Numeric	0.422^{***}
	(0.0004)
Constant	-10.246^{***}
	(0.015)
Observations	1,568,548
Log Likelihood	−5,832,379.000
Akaike Inf. Crit.	11,664,761.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 67: Age Estimation by Gender comparison for the CCD dataset with each gender group a reference label.

	Dependent variable:		
	diff_pred_age		
	(1)	(2)	(3)
GenderMale	−3.003***		4.686***
	(0.023)		(0.107)
GenderFemale		3.003***	7.689***
		(0.023)	(0.107)
GenderOther	−7.689***	−4.686***	
	(0.107)	(0.107)	
Constant	11.522***	8.519***	3.833***
	(0.016)	(0.017)	(0.106)
Observations	1,454,363	1,454,363	1,454,363
Log Likelihood	−5,895,877.000	−5,895,877.000	−5,895,877.000
Akaike Inf. Crit.	11,791,760.000	11,791,760.000	11,791,760.000
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 68: Age Estimation by Gender comparison for the UTKFace dataset with each gender group a reference label.

	Dependent variable:	
	diff_pred_age	
	(1)	(2)
genderMale	−0.428***	
	(0.021)	
genderFemale		0.428***
		(0.021)
Constant	3.939***	3.511***
	(0.015)	(0.015)
Observations	1,568,278	1,568,278
Log Likelihood	−6,278,200.000	−6,278,200.000
Akaike Inf. Crit.	12,556,403.000	12,556,403.000
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 69: Age Estimation by Skin Type comparison for the CCD dataset.

	Dependent variable:
	diff_pred_age
FitzDarker	-0.039^{*}
	(0.023)
Constant	10.018^{***}
	(0.017)
Observations	1,454,363
Log Likelihood	−5,905,835.000
Akaike Inf. Crit.	11,811,675.000
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 70: Age Estimation by Ethnicity comparison for the UTKFace dataset with each ethnicity group a reference label.

	Dependent variable:				
	diff_pred_age				
	(1)	(2)	(3)	(4)	(5)
ethnicityOther		-6.948^{***}	-4.242^{***}	-4.063^{***}	-2.797^{***}
		(0.041)	(0.044)	(0.045)	(0.045)
ethnicityWhite	6.948^{***}		2.707^{***}	2.885^{***}	4.151^{***}
	(0.041)		(0.029)	(0.031)	(0.030)
ethnicityBlack	4.242^{***}	-2.707^{***}		0.179^{***}	1.444^{***}
	(0.044)	(0.029)		(0.035)	(0.034)
ethnicityAsian	4.063^{***}	-2.885^{***}	-0.179^{***}		1.266^{***}
	(0.045)	(0.031)	(0.035)		(0.036)
ethnicityIndian	2.797^{***}	-4.151^{***}	-1.444^{***}	-1.266^{***}	
	(0.045)	(0.030)	(0.034)	(0.036)	
Constant	-0.944^{***}	6.004^{***}	3.298^{***}	3.119^{***}	1.854^{***}
	(0.037)	(0.017)	(0.024)	(0.026)	(0.024)
Observations	1,568,548	1,568,548	1,568,548	1,568,548	1,568,548
Log Likelihood	−6,259,319.000	−6,259,319.000	−6,259,319.000	−6,259,319.000	−6,259,319.000
Akaike Inf. Crit.	12,518,647.000	12,518,647.000	12,518,647.000	12,518,647.000	12,518,647.000

Note: *p<0.1; **p<0.05; ***p<0.01

Table 71: Age Estimation by Lighting comparison for the CCD dataset.

	Dependent variable:
	diff_pred_age
lightingDark	-3.586^{***}
	(0.025)
Constant	11.040^{***}
	(0.014)
Observations	1,454,363
Log Likelihood	−5,895,983.000
Akaike Inf. Crit.	11,791,969.000
Note:	*p<0.1; **p<0.05; ***p<0.01