DHS SCIENCE AND TECHNOLOGY

Revisiting the Fitzpatrick Scale and Face Photo-based Estimates of Skin Phenotypes

October 29, 2020

John Howard, Yevgeniy Sirotin, & Jerry Tipton

The Maryland Test Facility

Arun Vemury

Director

Biometric and Identity Technology Center

Science and Technology Directorate

Homeland Security

Science and Technology

Disclaimer

- This research was funded by the U.S. Department of Homeland Security, Science and Technology Directorate on contract number 70RSAT18CB0000034.

- This work was performed by a dedicated team of researchers at the Maryland Test Facility.

- The views presented here are those of the authors and do not represent those of the Department of Homeland Security, the U.S. Government, or their employers.

- The data used in this research was acquired under IRB protocol.

Introduction

- Significant recent focus on how the performance face recognition algorithms **varies across demographics**, including race, gender, and age.

- **Many** potential underlying causes:
  - algorithm architecture, training set composition,
  - training image properties, test image properties,
  - face properties, individual behavior.

- Race categories are **problematic** for gaining insight:
  - Race categories are culture specific.
  - Individuals within a race category can vary in properties.
  - How race labels are assigned can vary wildly between datasets.
    - (e.g. Based on classifier for RFW vs. mugshot records for MORPH)
    - Any assignment (human or machine) except self report is going to have biases and error rates.

Introduction

- Face **phenotypes** have been suggested as the remedy.
- Phenotypes are **observable** characteristics, i.e. physical appearance.
- The 2018 Gender Shades paper was the first (?) to encourage this [1].
  - Assigned a numeric Fitzpatrick Scale number to **images** of individuals
  - Images of parliamentarians from different countries from government websites

[1]: Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." Conference on fairness, accountability and transparency. 2018.

Study	Year	Domain	Face Skin Phenotype Measure	Finding
Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. (Buolamwini and Gebru)	2018	Gender classification	**Fitzpatrick skin-type (FST)** assessed from analyzed sample.	Images of women with FST IV-VI misclassified more than those with FST I-III
Understanding Unequal Gender Classification Accuracy from Face Images. (Muthukumar et al.)	2018	Gender classification	**Fitzpatrick skin-type (FST)** assessed from analyzed sample. Y values assessed from analyzed sample (YCrCb colorspace).	Manipulating face lightness does not affect gender classification accuracy.
An Experimental Evaluation of Covariates Effects on Unconstrained Face Verification (Lu et al.)	2018	Face recognition	Six custom skin tone groups assessed from analyzed sample in IJB-B and IJB-C datasets.	Improved biometric ROC curves for lighter versus darker tones.
Model Cards for Model Reporting (Mitchell, et al)	2018	General Machine Learning	**Fitzpatrick skin-type (FST)**	Model cards provide benchmarked evaluations in a variety of conditions e.g. .. Fitzpatrick skin types
Predictive inequity in object detection. (Wilson et al.)	2019	Pedestrian detection	**Fitzpatrick skin-type (FST)** assessed from analyzed sample.	Pedestrians with FST IV-VI more difficult to detect relative to FST I-III.
Demographic Effects in Facial Recognition and their Dependence on Image Acquisition: An Evaluation of Eleven Commercial Systems. (Cook et al.)	2019	Face recognition	Relative reflectance assessed from independent sample image.	Images from individuals with lower skin reflectance produce lower similarity scores on some cameras.
Issues Related to Face Recognition Accuracy Varying Based on Race and Skin Tone (Krishnapriya et al.)	2020	Face recognition	**Fitzpatrick skin-type (FST)** assessed by human review of analyzed sample.	Increased FMR for subjects classified as Black or African American not associated with FST.

Study	Year	Domain	Face Skin Phenotype Measure	Finding
Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. (Buolamwini a	2018	Gender classification	Fitzpatrick skin-type (FST) assessed from analyzed sample.	Images of women with FST IV-VI misclassified more than those with FST
Understanding Unequal Gen Accuracy from Face Images al.)				face lightness does not classification accuracy.
An Experimental Evaluation Effects on Unconstrained Fa (Lu et al.)				metric ROC curves for darker tones.
Model Cards for Model Repo al)				provide benchmarked a variety of conditions e.g. skin types
Predictive inequity in object et al.)				with FST IV-VI more difficult tive to FST I-III.
Demographic Effects in Faci and their Dependence on Im An Evaluation of Eleven Co Systems. (Cook et al.)				individuals with lower skin roduce lower similarity me cameras.
Issues Related to Face Rec Varying Based on Race and Skin Tone (Krishnapriya et al.)		recognition	review of analyzed sample.	IR for subjects classified as Black or African American not associated with FST.

"… the Fitzpatrick I–VI skin tone rating is the appropriate choice for this article due to its simplicity and widespread use, including prior use in the face recognition research community; e.g., metadata for face images in the IARPA IJB datasets [32], work by Buolamwini and Gebru [7], Lu *et al.* [30], and Muthukumar *et al.* [34]."

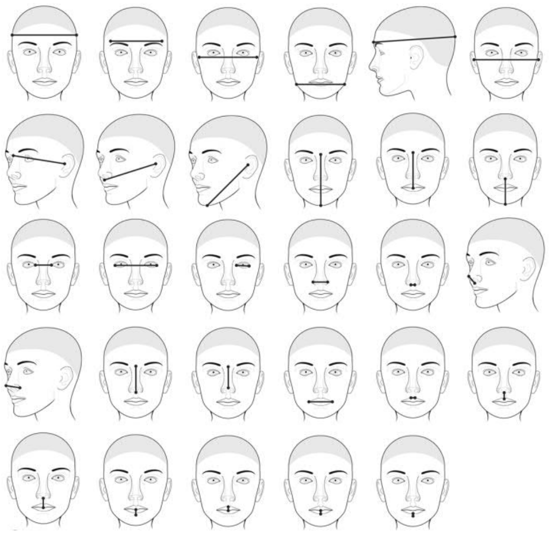
- Krishnapriya et al.

If we are reaching a consensus standard measure, is it the right one?

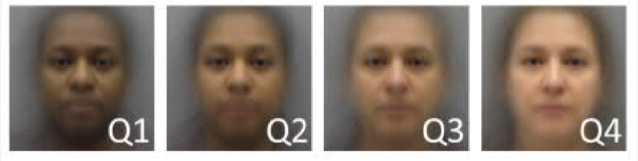
And are we measuring it the right way?

Face Phenotypes

Face Structure
(size, shape of face) [1]



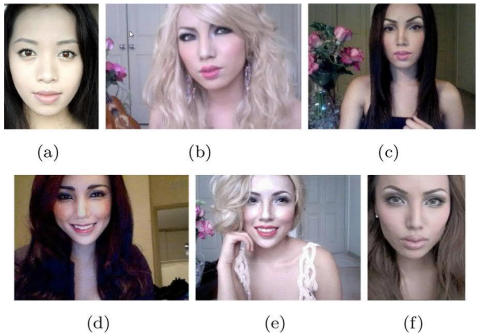
Face Skin Color
(melanin, hemoglobin, thickness)



We will focus on assessing one phenotype:

Face Area Lightness (FAL)

Face Styling
(tattoos, hairstyle, & makeup) [2]



[1]: Kesterke et al. “Using the 3D Facial Norms Database to investigate craniofacial sexual dimorphism in healthy children, adolescents, and adults” Biology of Sex Differences (2016) 7:23
[2]: Dantcheva, Antitza, C. Chen, and A. Ross. "Makeup challenges automated face recognition systems." *SPIE Newsroom* (2013): 1-4.

Face Area Lightness and Color Measures

Categorical:

- IARPA Janus Benchmark (IJB)
  - (1) light pink, (2) light yellow, (3) medium pink/brown, (4) medium yellow/brown, (5) medium dark brown, and (6) dark brown
  - Subjective based on human review of images
- Fitzpatrick Skin Type (FST)
  - (I) always burns, (II) burns easily, (III) sometimes burns, (IV) burns minimally, (V) rarely burns, and (VI) never burns*
  - Subjective self-report
  - Subjective expert assessment

* Only burning components of Fitzpatrick categories included for brevity

Continuous:

- Measured from face area on photographs or using calibrated instruments directly from face skin
- Individual Typology Angle (ITA)
  - Used in Diversity of Faces Dataset
  - Angle in the L* - b* color plane in the L*a*b* color space
- Face Area Lightness
  - Y in YCrCb color space
  - L* in L*a*b* color space
  - Y in XYZ color space

Data

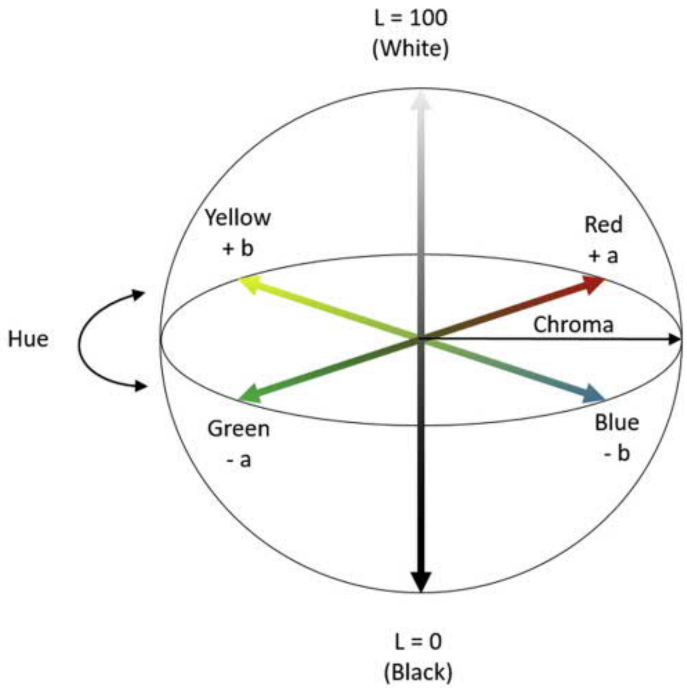
- Data collected during the 2019 Rally or publicly available
  - MEDS
  - Enrollment mages
  - Acquisition system images
  - Images from a “historic gallery”
  - Calibrated skin tone measurements

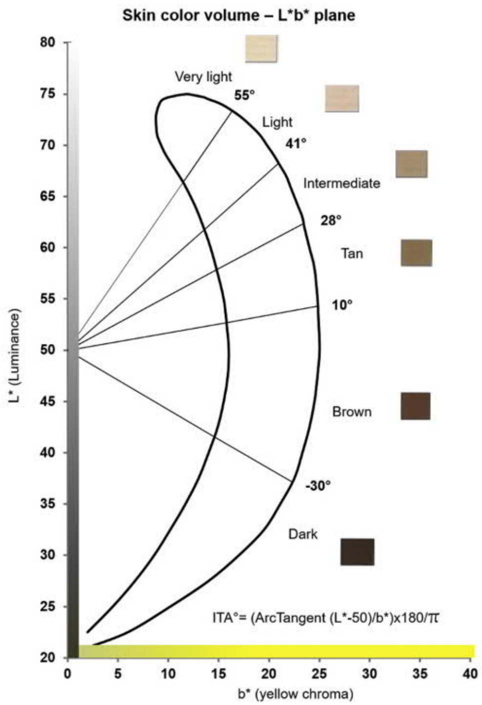
- Estimates of photo based skin tone were taken for each image and arranged into datasets that ranged from a varied environment, capture time, and device:

Dataset	Source	Environment (E)	Time (T)	Device (D)	Face Lightness Measure
MEDS	MEDS	Varied	Varied	Varied	L_f
CE	Historic & Acquisition	Constant	Varied	Varied	L_f
CET	Acquisition	Constant	Constant	Varied	L_f
CED	Historic & Acquisition	Constant	Varied	Controlled	$L_{f,d} - \mu_{f,d} + \mu_f$
CEDT	Acquisition	Constant	Constant	Controlled	$L_{f,d} - \mu_{f,d} + \mu_f$
Corrected	Enrollment	Constant	Constant	Constant	$(L_{f,d} - L_{b,d}) + \frac{1}{2}(\mu_{f,d} - \mu_{b,d})$
Calibrated	Colormeter	Constant	Constant	Constant	$\frac{1}{2}(L_{rc} + L_{lc})$

TABLE 3
Face lightness datasets examined.

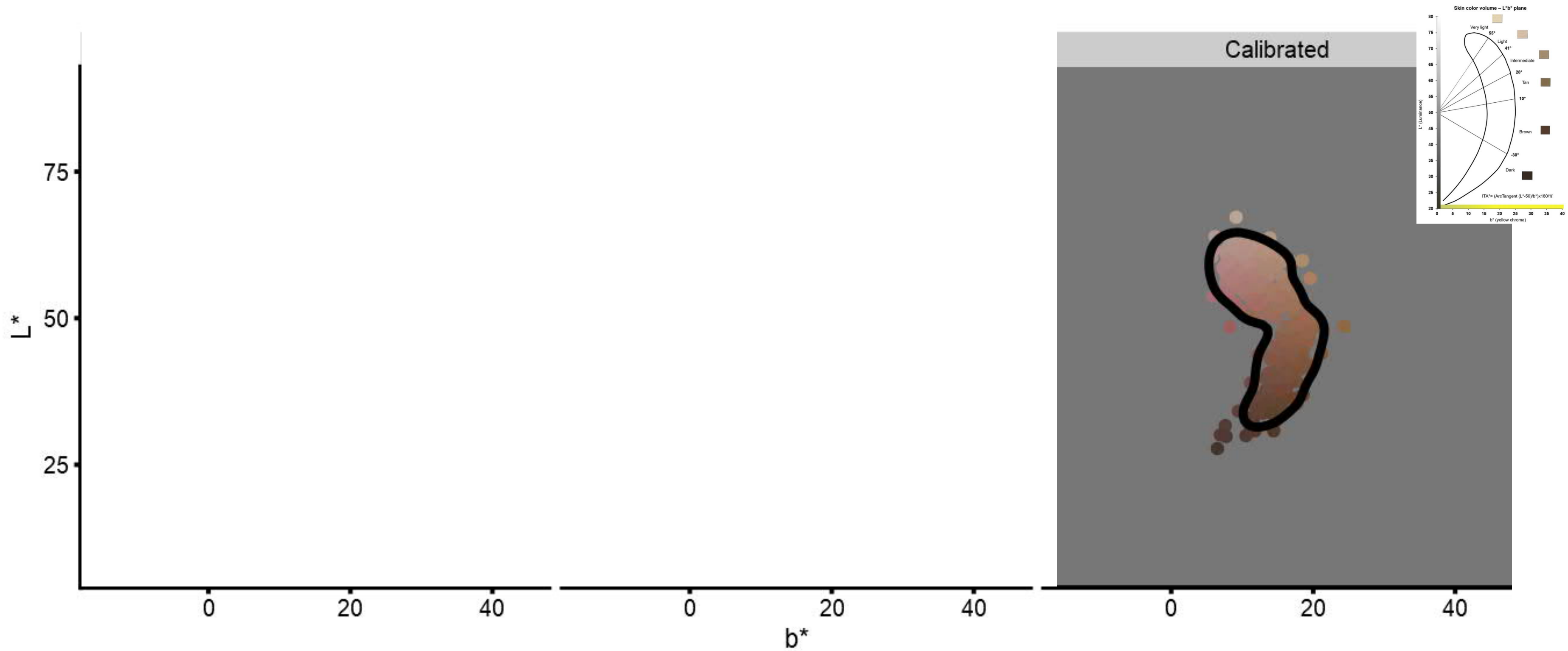
Face Area Lightness and Color Space



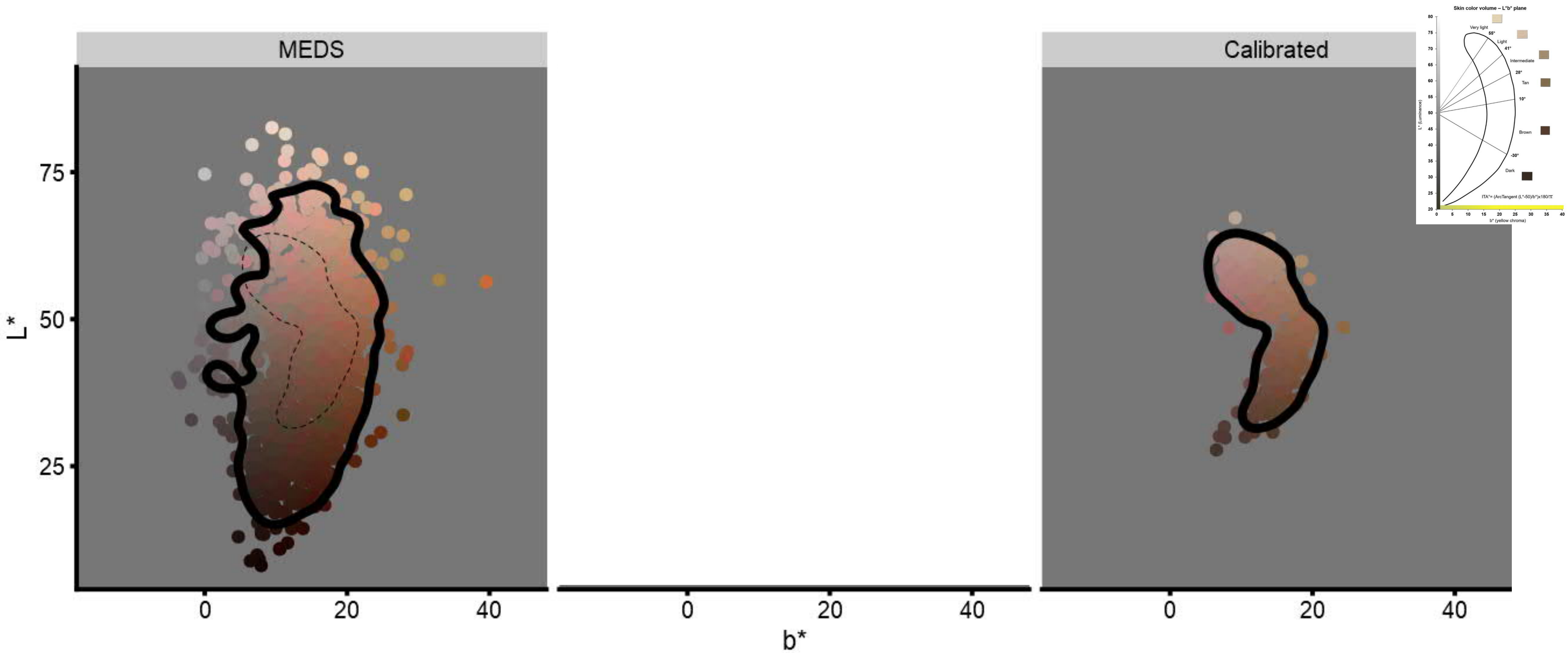


[1] Ly, B. C. K., Dyer, E. B., Feig, J. L., Chien, A. L., & Del Bino, S. (2020). Research techniques made simple: cutaneous colorimetry: a reliable technique for objective skin color measurement. Journal of Investigative Dermatology, 140(1), 3-12.

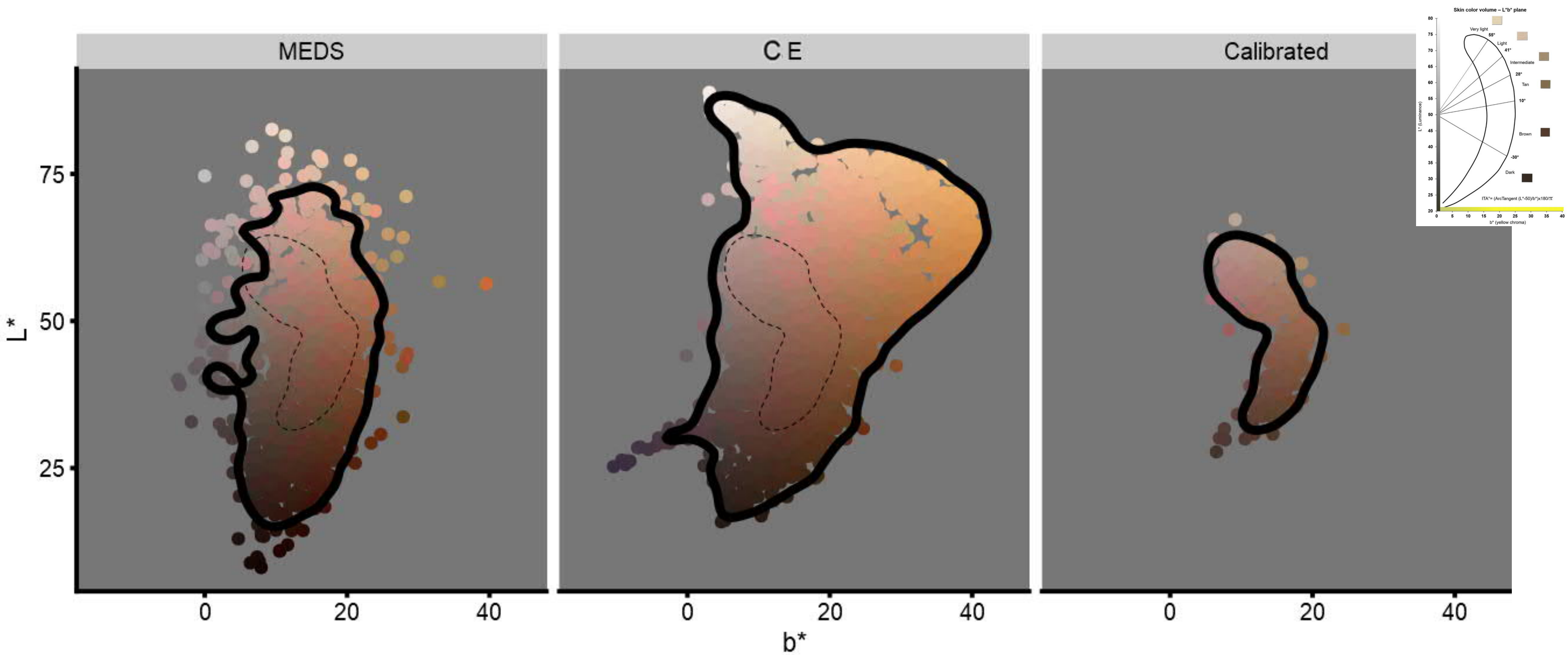
Control in Face Area Lightness and Color Measurement



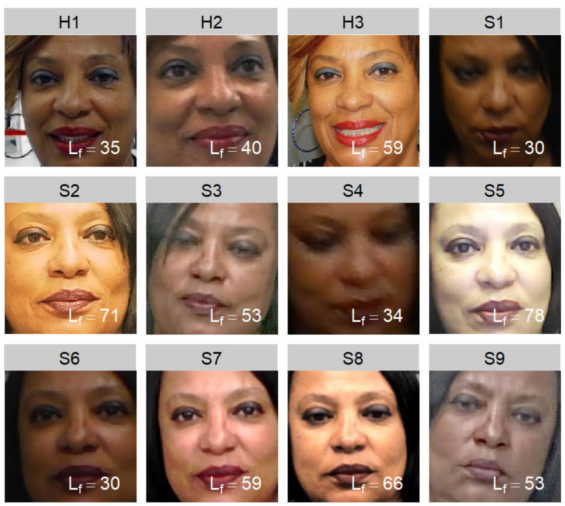
Control in Face Area Lightness and Color Measurement



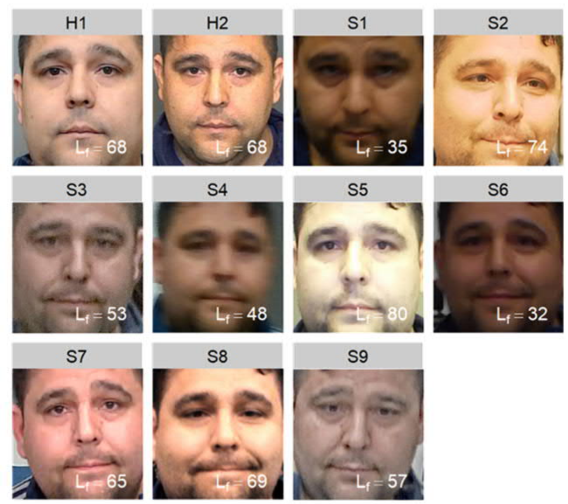
Control in Face Area Lightness and Color Measurement



Variation in Face Area Lightness Measurement



78-30 = 48

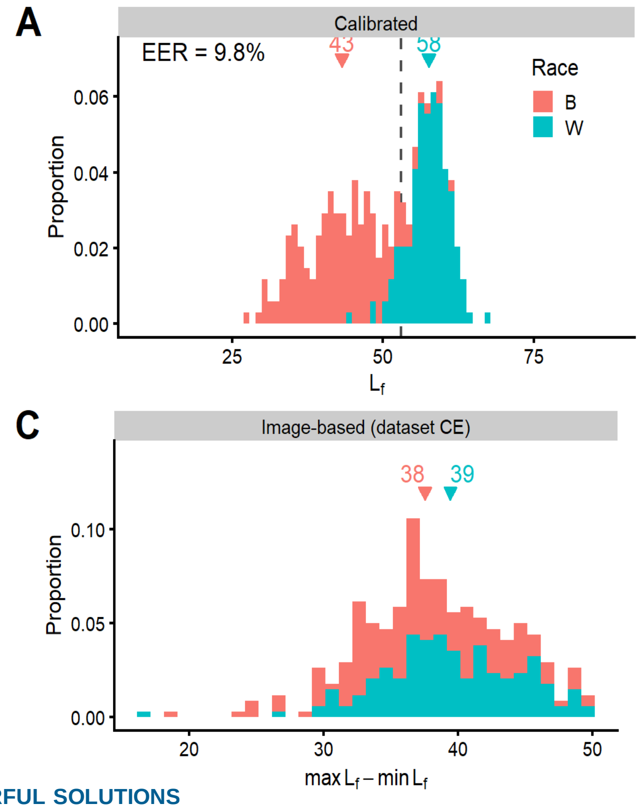


80-32 = 48

- Images of the same individual taken by different systems and times show more than 2-fold variation in Face Area Lightness

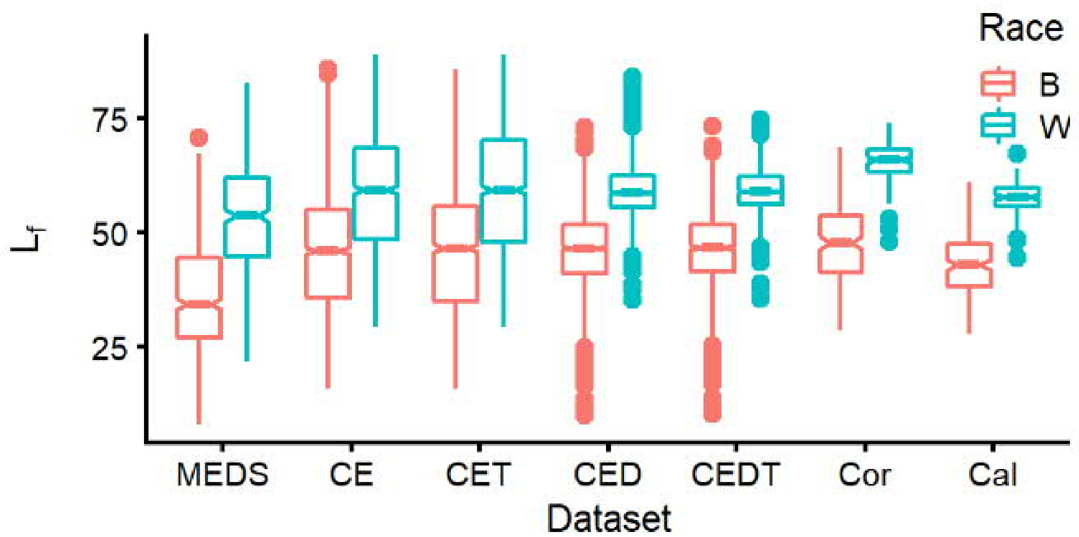
Variation in Face Area Lightness Measurement

- This variation for a single person (e.g. 48) is larger than most differences across demographic groups.
- In other words, the **error on the measurement** is larger than the measurement when using photo based estimates of skin tone.
- Error was consistent for both subjects who self identified as White and Black.



Variance in Face Area Lightness Measurements

- Better control in image acquisition generates better quality Face Area Lightness estimates:
  - **MEDS –** MEDS II
  - **CE** – controlled environment (MdTF)
  - **CET** – CE and controlled time (MdTF Rally 2 Images)
  - **CED**– CE and subtracting within-device mean
  - **CEDT** – CET and subtracting within-device mean
  - **Corrected** – enrollment images with background correction
  - **Calibrated** – DSM III color meter



No Control → Maximum Control

Rethinking Fitzpatrick

- FST is a questionnaire originally designed to determine the appropriate dose of oral methoxsalen for treating psoriasis using photochemotherapy in white individuals [1]
- FST is not skin color, in fact FST is known to be a **generally unreliable estimator of skin pigmentation**
- The FST was developed explicitly because dosing based on observed phenotypes (hair and eye color) led to medical error
- There is mounting evidence from the medical community that FST can be less reliable as an assessment for non-White individuals

Editorial

The Validity and Practicality of Sun-Reactive Skin Types I Through VI

The concept of sun-reactive "skin typing" was created in 1975[1] for a specific need: to be able to classify persons *with white skin* in order to select the correct initial doses of ultraviolet A (UVA) (in joules per cubic centimeter) in the application of the then newly developed technique for the treatment of psoriasis—oral methoxsalen photochemotherapy (PUVA).[2] The need arose as a result of experience with several patients who were a "dark" phenotype (brown or even black hair, and some with brown eyes) but, to our surprise, developed severe phototoxic reactions following oral ingestion of 0.6 mg/kg of methoxsalen and then, two hours later, were exposed to 4 to 6 J/cm². These initial doses were obviously too high, and it was then understood that the estimation of the white-skinned person's tolerance level to oral PUVA could not be based solely on the phenotype (hair and eye color). A simple approach was necessary for the impending large-scale oral PUVA photochemotherapy trials in the United States in the mid-1970s.[3,4] It was decided that a brief personal interview regarding the history of the person's sunburn and suntan experience was one approach to estimate the skin tolerance to ultraviolet radiation (UVR) exposure.

and a light tan at seven days." This group is skin type II. These are fair-skinned individuals with blond, red, or brown hair, green or hazel eyes, and skin that burns and peels easily. These individuals tan slightly only after repeated exposures. Also, a subgroup of skin type IV will respond: "A slightly tender burn at 24 hours and a moderate tan at seven days." This is skin type III and is the largest group in the United States.

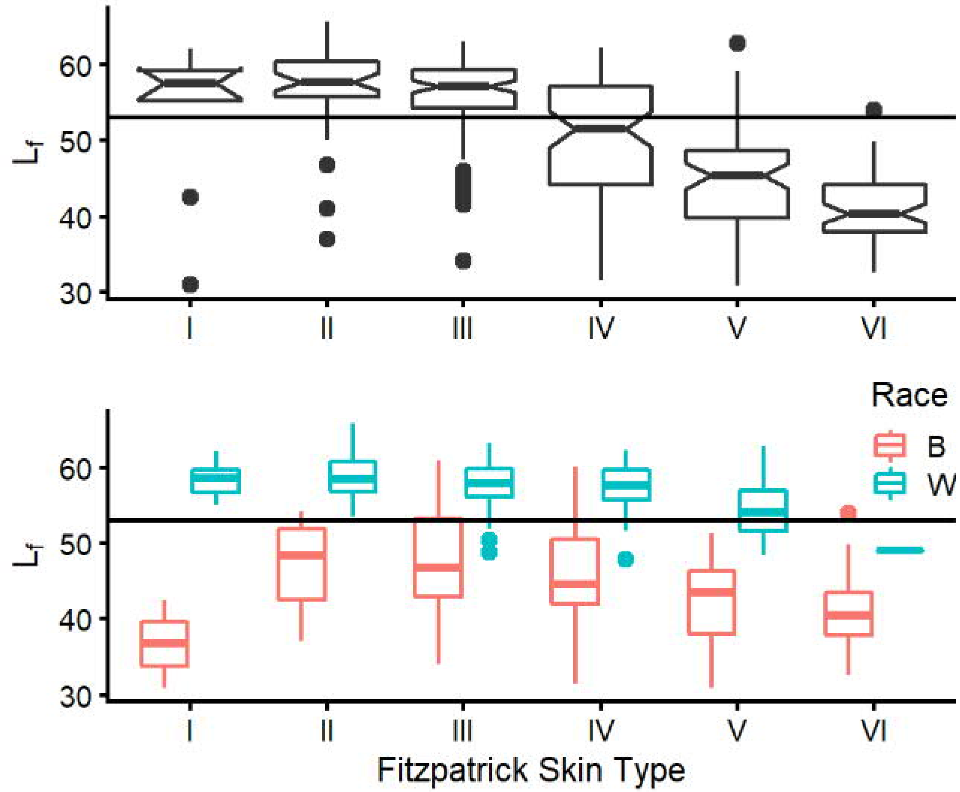
Individuals with skin type I have no inherent melanin pigmentation (ie, constitutive melanin pigmentation) and develop a marked tender sunburn or erythema following short exposures to UVR (sunlight or artificial ultraviolet B [UVB]) and are absolutely incapable of tanning (facultative melanin pigmentation). Persons with skin type I are keenly aware of their intolerance to sunlight and many give the same story: "I never go out in the direct sunlight, and when I did go out in my youth, I would only burn and peel. I have actually had severe blistering sunburns requiring bed rest for a couple of days. I never tan at all."

Persons with skin type IV, on the other hand, although exhibiting white skin with no clinical evidence of inherent melanin pigmentation, will usually

[1]: Fitzpatrick, T. B. (1988): The validity and practicality of sun-reactive skin types I through VI. In *Archives of dermatology* 124 (6), pp. 869–871. DOI: 10.1001/archderm.124.6.869.

FST (self reported) is not a Measure of Face Area Lightness

- We assessed Fitzpatrick scores based on self report
  - 363 volunteers taking part in the 2019 Biometric Technology Rally
- Compared with Face Area Lightness measured using a calibrated color meter
- Face Area Lightness decreased with higher FST
  - But this is because different proportions of volunteers of each race group chose each FST category
  - Face Area Lightness changed relatively little with FST within each race group

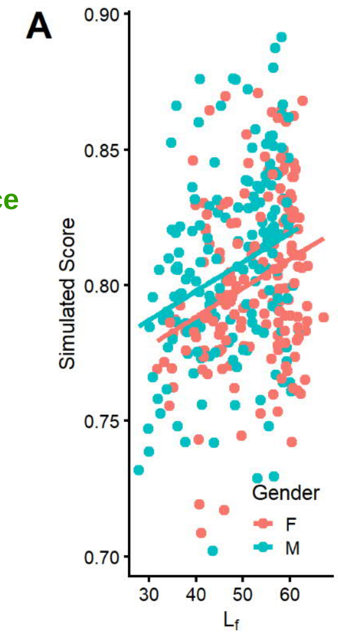


Poor Measures Cause Errors in Models of Biometric Performance

- Linear models are often used to statistically measure the effect of covariates on biometric performance, but they make errors:
  - **Type I error:** when an effect NOT really present is discovered
  - **Type II error:** when an effect really present is NOT discovered

- Consider a notional biometric system whose **Score** depends on Age, Gender, and **Face Area Lightness, but NOT Race**

$Score \sim f(Age, Gender, L_f)$ **NOT** $Score \sim f(Age, Gender, Race)$

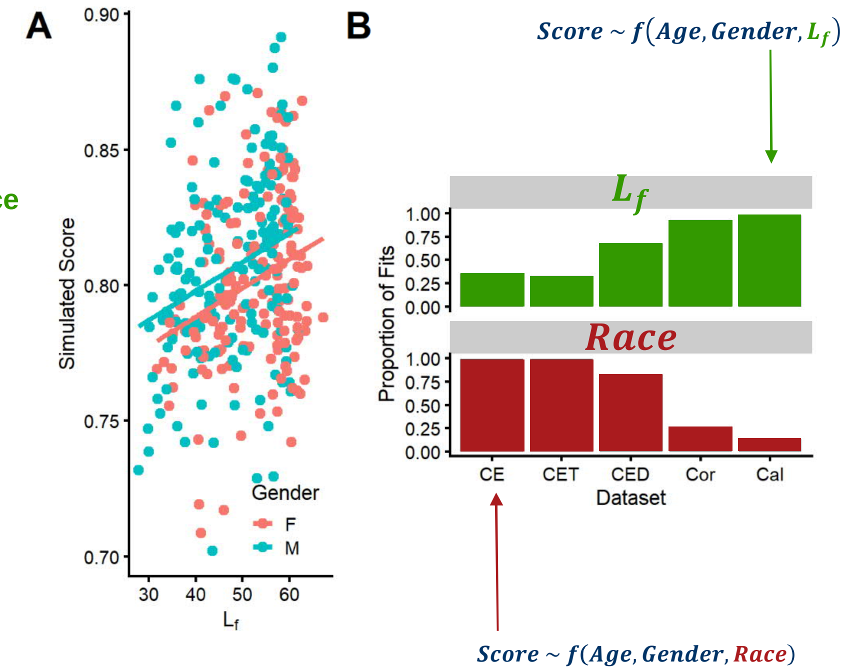


Poor Measures Cause Errors in Models of Biometric Performance

- Linear models are often used to statistically measure the effect of covariates on biometric performance, but they make errors:
  - **Type I error:** when an effect NOT really present is discovered
  - **Type II error:** when an effect really present is NOT discovered
- Consider a notional biometric system whose **Score** depends on Age, Gender, and **Face Area Lightness, but NOT Race**

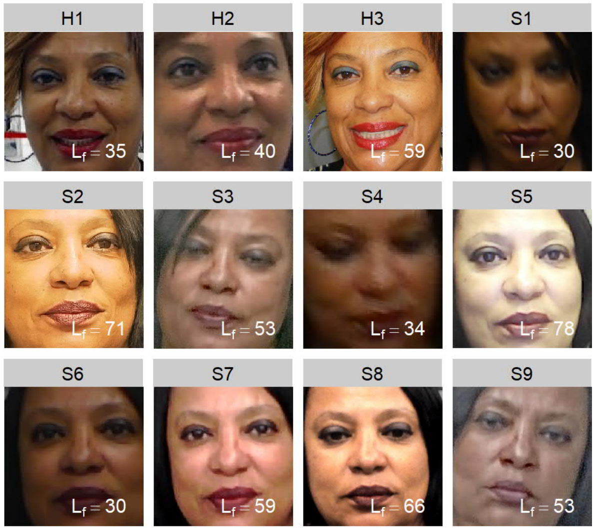
$$Score \sim f(Age, Gender, L_f) \quad \text{NOT} \quad Score \sim f(Age, Gender, Race)$$

- Simulated 1,000 random datasets gathered from this notional biometric system and fitted models to each dataset substituting different measures of Face Area Lightness
- Models using controlled measures had low error rates
- Models using **uncontrolled measures** of Face Area Lightness were **prone to error:**
  - **100% Type I error:** all models incorrectly found the effect of Race
  - **75% Type II error**: only ~25% of the models correctly found the effect of Face Area Lightness
- **This may happen in real models that use uncontrolled measures**



Conclusions

- The computer vision community recently began categorizing skin phenotypes in images using 6-point scales referred to as “Fitzpatrick Skin Types”
- Calling these measures FST is problematic for the following reasons:
  - As originally developed, FST is assessed by a survey to **measure sensitivity to UV light**
  - FST is measured by self report or by a physician direct assessment
- FST as originally defined is not an appropriate measure of skin color
  - FST has been shown in the medical literature to be an **unreliable estimator of skin pigmentation**
- All existing work applying FST to computer vision has involved human raters judging the skin pigmentation of subjects in images
  - 6-point skin tone classifications schemes have been conflated with FST
  - **These measures likely do not reflect FST**

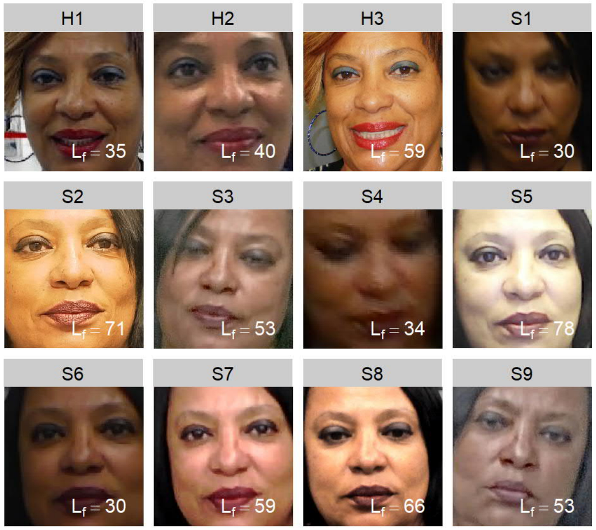


Conclusions

- Estimating skin phenotypes from uncontrolled images is subject to significant intra-subject variation
  - We show how lack of color control **affects automatic measures** of Face Area Lightness from images
  - Lack of color control is ***likely* to also confound human estimates** of Face Area Lightness from images

- To measure the relationship between biometric performance and phenotypes, we need controlled and careful measurement
  - Images scraped from the web often do not meet these criteria
  - We have shown how this **lack of control can lead to incorrect statistical conclusions**

- Measuring phenotypes correctly may require collection of **new samples**, but will **prevent errors** in statistical inference



Questions?

- This work was performed by a dedicated team of researchers at the Maryland Test Facility.
- Find out more at https://mdtf.org/
- john@mdtf.org
- yevgeniy@mdtf.org
- jerry@mdtf.org
- arun.vemury@hq.dhs.gov