Testing a Prenatal-Hormone Effect-Size Hierarchy in Facial Morphology: A Pre-Registered Self-Refutation at Celebrity Scale

Michael Abdo
Independent Researcher · Correspondence: Michael@michaelabdo.com

Pre-registered empirical report.

Abstract

A multi-axis common-cause model of prenatal hormone exposure predicts that a trait-axis's readability from facial morphology increases monotonically with the strength of its prenatal hormonal organization. Operationalized on a community-typed celebrity facial dataset, the model forbids a specific effect-size ordering: a "function" axis reads at least as strongly as a sex-modality axis, and both out-read two weakly-organized "letter" axes (function $\geq$ sex-modality $>$ letters). An exploratory discovery cohort (707 identities) appeared to support this - a function-axis effect of Cohen's d ≈ -0.47 surviving sex adjustment - but that analysis selected the maximum-effect principal component per axis, inflating small effects toward the predicted hierarchy. We therefore pre-registered an adversarial confirmatory test: a single frozen whitened-LDA estimator on the full 40-dimensional Basel Face Model shape vector, with identity-disjoint cross-validation and sex-residualization, timestamped on OSF with SHA-256 fingerprints. It refuted the hypothesis on 83 held-out identities; a powered replication of the byte-identical estimator (549 identities) refuted it again. The sex-residualized ordering was sex-modality ($|d| = 0.39$) > a letter axis (S_N, $|d| = 0.27$) > function ($|d| = 0.21$) > the other letter axis (F_T, $|d| = 0.12$): function ranked below a letter axis and failed FDR correction ($p = 0.17$). The one robust survivor, the sex-modality coin (d ≈ 0.39, p $= 1.1 \times 10^{-5}$), reads beyond the biological-sex direction - the classic physiognomy failure mode in which a face→trait signal reduces to a demographic shortcut. A well-powered test largely excludes the strong function claim, while a genuinely small function effect (d ≈ 0.20) remains undetermined and, given exhaustion of the rare lead-class, unanswerable with this population. The contribution is methodological: a pre-committed test that falsifies its own author's prior, and a demonstration that the surviving face→trait signal is a sex confound, not a developmental trace.

---

1. Introduction

Physiognomy, the practice of inferring character from facial appearance, collapsed as a science because it insisted on a uniform legibility of character that could not withstand confounds. A modern lineage of machine learning, driven by increasingly general learned image representations such as CLIP (Radford et al., 2021), has brought the empirical question back without the moral one. Face images have been reported to predict sexual orientation (Wang and Kosinski, 2018), Big Five personality traits (Kachur et al., 2020), and political orientation (Kosinski, 2021). A parallel literature reads facial metrics as proxies for hormonally-mediated behavior, most prominently the facial width-to-height ratio (fWHR) and aggression (Haselhuhn, Ormiston and Wong, 2015). Two cautionary findings discipline any such claim. First, Agüera y Arcas, Mitchell and Todorov (2018) showed that an in-the-wild facial signal can arise entirely from counterfeit sources like pose, grooming, makeup, camera angle, and sun exposure rather than from morphology. This means an apparent face-to-trait effect risks measuring self-presentation, not the body. Second, the

fWHR-to-aggression effect failed to replicate at scale. The 2015 meta-analysis pooled to a modest r of approximately .11 (d of approximately 0.22, based on our r-to-d conversion), with the effect reported as larger in laboratory than in field studies. Kosinski (2017), across 137,163 individuals, found fWHR unrelated to 55 psychometric constructs. The magnitude that individual lab studies in this subfield over-read, |d| of approximately 0.4 to 0.5, which is the range of our own discovery effect (d of approximately 0.47), is exactly the fragile range in which a selected, in-sample point estimate can masquerade as a robust effect.

1.1 The model and the falsifiable ordering

The motivating model, developed in the companion theory paper by Abdo (2026a), generalizes the organizational hypothesis. This hypothesis posits that perinatal hormones permanently shape both soma and brain. The model moves from single, clinically-extreme axes, such as congenital adrenal hyperplasia, to a multi-axis common cause. The premise that a prenatal-hormone signal can leave a measurable, sexually-dimorphic facial trace is independently grounded. Whitehouse et al. (2015) show that prenatal testosterone exposure predicts sexually dimorphic facial morphology in adulthood. The 2D:4D digit-ratio proxy for prenatal testosterone is widely used. It shows only weak and heterogeneous links to circulating testosterone across 54 studies (Sorokowski and Kowal, 2024). This serves as a caution that any single morphological readout of prenatal hormones is, at best, a small and noisy signal. Under this model, the intrauterine hormonal milieu jointly specifies a morphological map, including facial geometry, and a neural map, or dispositional architecture, from a shared developmental input. Observable form is therefore a correlated trace of the same event that shaped disposition. It is not a cause of it. Critically, this model departs from physiognomy by forbidding something specific. Define, for each trait-axis a, its hormonal organization O(a), which indicates how strongly axis a is specified by the developmental input. The model predicts:

> Readability of axis *a* from facial morphology increases monotonically in O(a).

This is a hierarchy prediction, not a point prediction, and it is falsifiable in three ways. A flat ordering across axes implies no common-cause coupling. An inverted ordering implies the model is wrong. And a weakly-organized axis that reads strongly implicates an alternative cause, such as demographic or self-presentation factors. No prior physiognomic claim made an ordering prediction. They claimed uniform legibility. This is precisely what confounds counterfeit, because image-quality and demographic artifacts do not respect an a-priori O(a) tiering derived independently of facial data.

When operationalized through the Objective Personality System (OPS) coin taxonomy, prior literature ranks the axes with sexually-dimorphic or sex-modality axes (high O) at the top, followed by function axes (intermediate O), and finally "letter" axes (low O). Before examining any held-out facial data, the strict confirmatory prediction was set as follows:

> **function $\geq$ sex-modality $>$ letters (S_N, F_T)**, with the function and sex-modality coins each separable above chance and the letter coins at or near the null.

1.2 Why we tested it adversarially

An exploratory discovery analysis appeared to support this ordering: on a 707-identity discovery cohort ($\geq$3 fitted photos each, aggregated to per-identity 3D shape centroids), a function-axis coin (observer/decider, Se vs Te) showed a facial effect of Cohen's d ≈ -0.47 that survived adjustment for biological sex (sex-residualized d ≈ -0.45), recurred within each sex separately, and survived

within-ethnicity adjustment. But this discovery analysis selected, per coin, the principal component (of the first five) that maximized the effect. Selecting the winning component inflates small effects and biases the ordering toward whatever hierarchy is hoped for; in the $|d| \approx 0.4$-0.5 regime this is exactly the inflation that individual fWHR lab studies over-read (lab estimates ran larger than field before the meta-analytic pooled effect fell to $d \approx 0.22$, our $r \to d$ conversion). We therefore declined to treat the discovery result as confirmation.

We adopted an adversarial protocol to test whether our prior assumption was false. First, we committed to a single estimator that utilizes the entire shape vector without any component selection. Second, we defined a held-out cohort of identities that remained untouched during the discovery analysis. Third, we established a directional decision rule specifying which outcomes, such as flat ordering, inverted ordering, or a non-separable function coin, would constitute refutation. Fourth, we pre-committed to neither substituting discovery point estimates to rescue a failed test nor re-selecting the estimator after observing held-out effects. Fifth, we timestamped the entire plan on OSF using SHA-256 fingerprints to freeze the estimator code and cohort. This process allows any later alteration of the estimator or cohort after registration to be falsified by re-hashing. This step removes the flexibility of estimator and cohort selection rather than proving that effect sizes were computed only afterward, which represents the limit of the hash chain as made precise in §2.3. We then ran the test. The first run refuted the hypothesis but left the rarest class underpowered. We subsequently registered and ran a second, larger test using the byte-identical estimator. This paper reports both tests.

---

2. Methods

2.1 Data and cohorts

Each analysis unit represents a single public figure, or celebrity, who carries a community-assigned OPS type. We collected photographs for each named identity from publicly available web sources, organizing them into one folder per person. We did not use private or non-public images, and we release only derived 3D shape features rather than the raw photographs (see data-provenance and human-subjects basis in §6). A folder entered the manifest only if its type string parsed to a complete 5-coin OPS stack. Per-identity 3D fits then required at least three successfully-fitted photos. We de-duplicated identities by name to ensure no person is double-counted. For each identity, we aggregated the photos to a single per-identity 40-dimensional shape centroid, which is the mean over that person's fitted photos. This process yields one analysis row per identity. Cross-validation is therefore identity-disjoint by construction, and no photo-level leakage arises.

Three cohorts entered this work. Each held-out set was disjoint from the discovery cohort. The smaller held-out set is contained in the larger, as detailed below.

- **Discovery (exploratory only; 707 identities).** The 707-identity discovery cohort (each with $\geq$3 fitted photos) carries the selected-component discovery effects. It plays no role in either confirmatory test except to define which identities are *excluded* from the held-out sets - the held-out cohorts are defined as identities absent from the 707-identity discovery manifest.
- **Held-out, registration 1 (83 identities).** Every OPS-typed identity with a full 5-coin type and $\geq$3 fitted photos absent from the 707-identity discovery manifest, as frozen at the first registration (1,364 photos, 100% fitted; 81 of 83 matched a biological-sex label). The lead function-coin split here was small: 10 lead-Se vs 12 lead-Te.

- **Held-out, registration 2 / powered (549 identities).** The complete set of OPS-typed identities with a full 5-coin type and ≥3 fitted photos absent from the discovery cohort (9,966 photos, 100% fitted; 547 of 549 sex-matched; the same two identities, Ashley Banfield and Cody Talks, lacked a sex label in both runs). Lead function-coin split: 162 lead-Se vs 59 lead-Te.

The pool of typed-and-photographed identities grew between the two registration points because more identities were typed, photographed, and 3D-fitted during that interval. Consequently, "every" and "complete set" each denote completeness as of that specific cohort's own freeze date. The 83-cohort represents every qualifying held-out identity as frozen at registration 1, while the 549-cohort comprises every qualifying held-out identity as frozen at registration 2. Since the pool only expanded between these freezes, with identities added but none dropped or relabeled under identical inclusion criteria (full 5-coin OPS type, ≥3 fitted photos, and absent from the 707-identity discovery cohort), the 83 are contained within the 549 rather than forming a separate population.

2.2 The frozen estimator

The pre-registered estimator logic is identical across both registrations because the run executes the same whitened-LDA helper path in both. Yet the artifact fingerprinted at each registration differs in file name. Registration 1 fingerprinted `confirmatory_analysis.py` (SHA-256 `578ec8f4…`), which the run executes unchanged via the `confirmatory_heldout.py` wrapper. Registration 2 fingerprinted the run wrapper `confirmatory_heldout.py` (SHA-256 `c22dcb076658bca3729c05554dc2c6dbe6dab0a56384111e86485fa42a548df3`), executed byte-identically to the first run. The estimator specification, common to both, is:

- **Features.** The full 40-dimensional Basel Face Model (BFM; Paysan et al., 2009) shape coefficient vector per identity, obtained by fitting each photo with 3DDFA_V2 (3D Dense Face Alignment; Guo et al., 2020), which regresses a 62-dimensional 3D morphable-model parameter vector per face. We retain only the 40-dim shape block (parameter indices 12-51); pose (indices 0-11) and expression (52-61) are discarded from the feature set. Per-identity shape centroids are the analysis units. **No principal-component selection** - the whole shape vector is consumed, eliminating the per-coin component choice that made the discovery analysis exploratory.
- **Separability.** Per coin, a covariance-whitened linear discriminant (equivalently, the cross-validated Mahalanobis separation between group centroids) expressed as a Cohen's-d-equivalent via the held-out AUC.
- **Cross-validation.** Identity-disjoint GroupKFold (k = 5) on person identity. One row per identity makes folds leakage-free by construction.
- **Primary adjustment (sex-residualization).** The shape vector is projected onto the null space of the linear biological-sex direction before computing the discriminant. Biological sex is read from the annotation database (`persons.gender`), never inferred from the image - the specific defense against the apparatus simply re-reading sex from facial appearance.
- **Confound covariate (pose).** Head pose (yaw, pitch, roll) recovered from the 3DDFA_V2 pose block is regressed out, and the function effect re-estimated after sex+pose adjustment versus sex-only. (Registration 1, §3, additionally listed a FairFace ethnicity covariate among the frozen confounds; the frozen confirmatory code never wired it into the analysis path, so it was not executed on either held-out run - a registered-vs-executed gap detailed in §5.)
- **Uncertainty.** Bootstrap 95% confidence intervals over 1,000 identity resamples per coin.
- **Multiplicity.** Benjamini-Hochberg false-discovery-rate correction (q = 0.05) across the four

coin tests.

The decision rule was directional and pre-committed. The §2 ordering holding on held-out data with function and sex-modality both surviving FDR confirms the prediction. A flat or inverted ordering, or a function coin that fails to separate above chance after sex+pose adjustment, refutes it. The held-out result was to be reported verbatim regardless of direction.

2.3 The two OSF registrations

Item	Registration 1	Registration 2 (powered)
OSF ID	osf.io/smrxt	osf.io/kj2rw
DOI	10.17605/OSF.IO/SMRXT	10.17605/OSF.IO/KJ2RW
Registration date	2026-06-15	2026-06-15
Confirmatory run date	2026-06-15 (post-registration)	2026-06-15 (post-registration)
Held-out n	83	549
Estimator SHA-256 (fingerprinted artifact)	`confirmatory_analysis.py`, `578ec8f4…` (frozen estimator; run via the `confirmatory_heldout.py` wrapper)	`confirmatory_heldout.py`, `c22dcb07…` (run byte-identical)
Frozen cohort SHA-256	`heldout_cohort_frozen.csv`, `ab664838…`	`heldout_large_cohort_frozen.csv`, `e5c5786f…`
Frozen features SHA-256	`heldout_features.npz`, `da019428…`	`heldout_large_features.npz`, `2002e174…`

The sequence register-1 to run-1 to register-2 to run-2 unfolded within a single calendar day (2026-06-15). In each instance, the OSF registration was timestamped first, and the confirmatory run followed later that same day, as each run log records "post-registration." We do not ask the reader to take the same-day ordering on trust. What the hash chain guarantees is mechanical, and we state its limit precisely. Each registration publishes, before its run, the SHA-256 fingerprints of the estimator code, the frozen held-out cohort, and the blind feature arrays (table above). Re-hashing those artifacts against the timestamped deposit proves that the estimator, the cohort, and the features were frozen and immutable at registration. This forecloses the degrees of freedom that matter most for a "forking paths" critique: post-hoc estimator re-selection, cohort swapping, and substitution of the discovery numbers. None of these actions can be done without changing a fingerprinted artifact. It does *not*, by itself, prove output-blindness. Because the estimator is deterministic on the already-frozen features, the effect sizes are a fully determined function of inputs that existed at registration time. Thus, re-hashing cannot exclude the possibility that the author ran that deterministic pipeline and observed its outputs before registering. The credibility of this test therefore rests on removed analytic flexibility. There was no estimator or cohort left to choose after the freeze, rather than on a provable temporal blindness of the effect sizes. We disclose, accordingly, that hypothesis-blind 3D feature extraction (which carries no hypothesis test) was permitted before registration. No per-coin effect size, Cohen's d, discriminant, or hypothesis test was performed by the author at registration time. Registration 2 is a powered replication of registration 1. The estimator, hypothesis, and decision rule are unchanged. The only difference is the larger frozen cohort, tested once, with no optional stopping. The first run's verdict stands as reported. The second is additive, not a replacement.

2.4 Sensitivity and the rare-class power ceiling

We calculated several metrics for the powered cohort (n = 549) to separate a true null result from one lacking sufficient statistical power. These calculations included the minimum detectable effect (MDE) per coin, the power to detect the effect claimed in the discovery, the 95% upper bound on the true effect, and the sample size a small function effect would necessitate. Here, power is measured as the two-sample Cohen's d at each coin's observed group split.

Despite the large overall sample size, the function coin is the least-powered coin because its minority arm, which consists of 59 lead-Te identities, caps the power. Its minimum detectable effect (MDE) at 80% power ($\alpha = 0.05$) is d = 0.43, whereas the three coins with balanced splits of approximately 250 versus 300 have an MDE of approximately 0.24. The power to detect the discovery-claimed function effect (d = 0.45) was 83%, but the power to detect a small effect (d = 0.20) was only 26%. The data place a 95% upper ceiling on the true function effect at d $\approx$ 0.60. Detecting a small d = 0.20 function effect at 80% power while keeping the natural 2.7:1 Se:Te ratio would require approximately 269 lead-Te identities, which is about 210 more than the 59 in hand. The typed label pool contains only 186 distinct lead-Te identities in total and only 14 unused ones. Collecting photographs for every unused lead-Te identity reaches a maximum cohort of 73, which is far short of 269. The small-effect function question is therefore not resolvable by collecting more photographs of the people already typed. It would require typing and photographing on the order of plus 83 to plus 291 brand-new lead-Te identities. This bounds, quantitatively, the only path that could rescue a weak function effect and shows that path is closed with the current typed population.

2.5 Typing-label provenance and the construct caveat

Each axis label originates from a per-identity OPS type stored in `persons.type_full`. We disclose three provenance facts and three caveats. *Facts:* (i) a two-rater agreement gate - a label persists only after two independent human typers concur, giving a structural reliability floor; (ii) labels are human-assigned from behavioral video, not model-assigned, so they do not leak from the facial encoder under test; (iii) typers assign coins from speech and behavior, keeping the label channel orthogonal to the facial-measurement channel. *Caveats, foregrounded:* (i) **OPS coins are not a validated psychometric instrument** - they carry no published construct- or criterion-validity coefficients, no test-retest, and no standardization against an external criterion; we treat each coin as a measured construct under test, and report effect sizes as the face-readability of *that construct*, not of any validated trait. (ii) The agreement gate discards the pre-agreement vote distribution, so a chance-corrected κ cannot be recovered from the stored records. (iii) Typers also view faces in the source video, so a face→label leakage path cannot be excluded a priori; this would, if anything, *inflate* a face→coin effect, which makes a null result conservative. Because of these caveats the contribution of this paper is the *method* - the pre-registered, adversarial test and its outcome - not the typology.

---

3. Results

3.1 Both tests refute the predicted ordering

The pre-registered confirmatory tests both rejected the hypothesis, adhering to their own frozen decision rules.

Registration 1 (n = 83). On the whole-shape sex-residualized estimator, the observed ordering was

function > letters > sex-modality. This represented an inverted ordering where the sex-modality coin fell to last, near the null. No coin survived BH-FDR correction because the smallest p value was 0.027, which is above the BH critical value of 0.0125 for that rank. Consequently, the confirmation condition failed since both function and sex-modality did not survive FDR. The function coin's large apparent |d| reflected a below-chance held-out AUC of 0.217. This means the held-out discriminant separated the classes in the opposite direction to discovery on a tiny split of 10 vs 12. Its pose-adjusted whole-shape effect was exactly +0.000. Per the §2.2 decision rule, this is a refutation. The small split made it an underpowered one. That is why we registered the powered replication.

Registration 2 (n = 549, powered). The function coin's split rose to 162 vs 59, which was well-powered against a large effect. The observed sex-residualized whole-shape ordering was:

Rank	Coin	sex-residualized |d|	held-out AUC	group n	raw p	FDR
1	sensory_mf (sex-modality)	**0.392**	0.609	304 / 243	1.1×10^{-5}	**survives**
2	S_N (letter)	0.266	0.574	283 / 264	2.6×10^{-3}	**survives**
3	o_or_d_1 (function)	0.214	0.560	162 / 59	0.172	fails
4	F_T (letter)	0.122	0.534	311 / 236	0.169	fails

The predicted ordering was function ≥ sex-modality > letters. The observed ordering was sex-modality > S_N (letter) > function > F_T (letter). Three independent components of the prediction failed simultaneously: (i) the function coin ranked *below* the sex-modality coin, contradicting `function≥{=latex} sex-modality`; (ii) the function coin ranked *below* a letter coin (S_N), contradicting `function > letters`; and (iii) the letter coins were not at the null - S_N (|d| = 0.27) exceeded the function coin and survived FDR, contradicting `letters at/near null`. The function coin failed FDR (p = 0.17), its bootstrap interval spanned zero (bootstrap d = +0.13, 95% CI [−0.38, +0.60]; the whitened-LDA point estimate is +0.21, AUC 0.560), and after sex+pose adjustment its whole-shape effect fell to d = +0.152 (CI spanning zero), so it did not separate above chance in the §2.2 decision-rule sense. Each of these is a named refutation condition in the frozen decision rule. The verdict is **REFUTES**, reported verbatim.

3.2 The discovery effect did not replicate; the n = 83 inversion was a power artifact

The discovery analysis reported a function effect of d ≈ −0.47 derived from a selected principal component. When applied out of sample to the whole shape vector without component selection, the function effect collapsed to +0.21. The inverted, below-chance reading observed at n = 83 (AUC 0.217) resolved, at n = 549, into a small positive but non-significant effect. Together, the two held-out runs show the discovery's magnitude did not survive. The selected-component point estimate was the inflation that the adversarial design was built to catch.

3.3 The robust survivor is a sex confound (the demographic-shortcut finding)

The OPS sex-modality coin (sensory_mf) emerged as the most resilient token throughout the powered test. Initially, at n = 83, it appeared near-null with a sex-residualized d of +0.392, an AUC of 0.609, an FDR p of 1.1×10^{-5}, and a bootstrap 95% CI of [+0.003, +0.604] that excluded zero. However, as the sample size expanded to n = 549, this token became the strongest and most robust effect. This reversal was driven by power, evident in the shift of its sex-residualized split from 40/41 to 304/243. By both name and construction, the sex-modality coin represents the axis most closely tied to biological sex. Our adjustment projects out only the linear biological-sex direction, leaving nonlinear or higher-order sex-linked structure in facial shape untouched. Consequently, a residual that survives on this axis is most parsimoniously interpreted as residual sex-linked demographic signal rather than as a sex-independent construct trace. This is why confound remains the best explanation despite residualization. It is the expected result for an apparatus that picks up a sex-correlated demographic signal that faces read this coin beyond the residualized biological-sex direction with a d of approximately 0.39. This outcome highlights the classic physiognomy failure mode: the surviving face-to-trait signal is overwhelmingly a demographic shortcut, not a developmental trace of a sex-independent disposition. Another token, the letter coin (S_N), also survived FDR with a d of 0.27. Since the model predicts this token to lie at the null, its survival counts against the model.

3.4 Sensitivity: the strong claim is largely excluded; a weak effect is unanswerable here

The null function is well-powered against the discovery's own claim. It boasts 83% power to detect d = 0.45 alongside a non-detection, with a 95% ceiling at d ≈ 0.60. Consequently, the discovery's magnitude is largely excluded. Although 0.45 lies inside the confidence interval, the design would more often than not have flagged it, and the discovery's central estimate did not replicate. The null function is, however, underpowered against a genuinely small effect. Power at d = 0.20 was only 26%. The observed point estimate of +0.21 sex-residualized and +0.15 pose-adjusted sits in this undetermined band. The CI half-width of 0.49 is roughly equal to the 80%-power MDE of 0.43, which is the signature of a test well-powered for a large effect and underpowered for a small one. This is no large effect and small effect undetermined, not a tight null around zero. Crucially, resolving the small-effect question is blocked by the rare-class ceiling of §2.4. The minority lead-Te arm is the binding constraint and the typed population is effectively exhausted, leaving 14 unused identities. Even the optimistic small-effect scenario cannot be reached without a major new typing-and-photography effort. Furthermore, such an effort would not restore the predicted hierarchy, since a function effect of ≈ 0.20 still sits below the sex-modality coin's 0.39.

---

4. Discussion

Three things follow from a pre-registered self-refutation.

The strong predicted hierarchy is refuted by these data, and the refutation is credible because it was hard to fake. The model forbade a specific ordering. The ordering was registered before the held-out data were touched, with the estimator and cohort bound by published SHA-256 fingerprints, and a directional decision rule that names refutation conditions in advance. Under that rule, two held-out tests rejected the prediction, once underpowered, once well-powered. A pre-registered null of this kind is worth more than an unregistered confirmation, because it removes the very

degrees of freedom, component selection, post-hoc estimator choice, selective reporting, that make in-the-wild face to trait effects so frequently irreproducible (cf. the fWHR to aggression reversal). The discovery effect ($d \approx -0.47$), obtained by selecting the maximum-effect component per coin, is a textbook instance of the inflation in the $|d| \approx 0.4$-0.5 range that the face-reading subfield has repeatedly over-read (individual fWHR lab estimates ran larger than field before the meta-analytic pooled effect fell to $d \approx 0.22$, our r to d conversion); the value of the present design is that it caught its own author's version of that error.

The surviving signal functions as a demographic shortcut rather than a developmental trace. Only one effect withstood the false discovery rate correction, and it targeted the sex-modality coin, which is the axis most tightly linked to biological sex. This result aligns precisely with how an apparatus that has latched onto a sex-correlated demographic signal would behave when read beyond the residualized sex direction. This scenario makes explicit the canonical physiognomy failure mode: when a face-to-trait pipeline appears to work, the mechanism driving that success is usually a demographic confound. Consequently, we do not claim that the face contains a sex-independent dispositional signal. The function axis, which was supposed to represent the strongest sex-independent trace, failed to separate from chance levels once biological sex and pose were removed.

The value of a registered null. It defines the question with precision (§3.4). This approach is uniquely useful for those tempted to just collect more data because the sensitivity analysis shows that path is closed. The rare lead-Te class is exhausted in the typed pool so no amount of additional photography of already-typed people can resolve the small-effect question. Even if a small effect were eventually established it would not restore the model's predicted ordering. The honest scientific output, given the data in hand, is a falsified strong hierarchy. It provides a quantified ceiling on the residual effect and a clearly-bounded open question that the existing population cannot answer.

---

5. Limitations

- **The construct is unvalidated; foreground the method, not the typology.** OPS coins carry no published construct- or criterion-validity coefficients, no test-retest, and no standardization against an external criterion. This paper's claims are about the *method* (a pre-registered, adversarial face→construct test and its outcome), not about the OPS typology, whose validity remains unestablished. Because typers also view faces in the source material, a face→label leakage path cannot be excluded; such leakage would inflate, never deflate, a face→coin effect, so the refutation is conservative with respect to it.
- **Single celebrity dataset.** All cohorts are public figures, and faces and fame are not independent (selection bias). A single dataset cannot establish generality; the demographic-shortcut finding in particular should be checked on non-celebrity samples.
- **No age covariate.** No age field was available for the identities, so age - a known driver of facial shape - could not be adjusted for. This is a declared limitation, not an adjustable covariate.
- **Ethnicity covariate not applied on held-out data.** The frozen confirmatory estimator did not wire in the FairFace (Kärkkäinen and Joo, 2021) ethnicity covariate (it was never coded into the confirmatory path), so ethnicity adjustment was omitted on both held-out runs. On the powered $n = 549$ run cached ethnicity labels existed for most identities (466/549), but the covariate was still not applied; the $n = 83$ run had no cached held-out ethnicity labels

at all (0/83). The sex-residualized primary analysis - the registered headline - is unaffected either way. We did not scrape new labels (the registration forbade it).

- **The small-effect function question is unanswerable with this population.** As quantified in §2.4 and §3.4, the binding rare-class ceiling means a genuinely small function effect can be neither confirmed nor excluded without a major new typing-and-photography effort targeting the minority lead-Te class.
- **Pose covariate fifth dimension.** The frozen estimator's fifth pose-related covariate used a within-identity pose-dispersion proxy rather than a true per-fit reconstruction residual (3DDFA_V2 does not store the latter); the actual yaw/pitch/roll pose adjustment is exact. This does not affect the verdict, which the function effect failed before that covariate was added.

---

6. Ethics statement

This paper takes a skeptical stance. Its central empirical finding is that an apparent face-to-trait signal, tested adversarially, refutes the developmental-ordering hypothesis that motivated the study. The one signal that survives this scrutiny is overwhelmingly a biological-sex confound, which serves as a demographic shortcut and represents the historical failure mode of physiognomy. We make no individual-prediction claim of any kind. Nothing in this work licenses screening, ranking, gating, or judging individuals, and we explicitly disclaim such use. We report a population-level, mechanism-level refutation, framed throughout against the misuse to which face-to-trait inference has historically been put. By design, the strongest single result of this paper, that a face-to-trait effect reduces to a sex confound, is a warning, not a tool. It is evidence that face-based trait inference detects demographics, not character.

Data provenance and human-subjects basis. The analysis units consist entirely of public figures. We obtained all photographs from publicly available web sources (§2.1). We collected no private, restricted, or otherwise non-public images. Nor did we use any identifying data beyond what each subject has already made public. As an independent researcher with no institutional affiliation, no IRB review was sought or available. The study involves no intervention, interaction, or contact with the individuals. It analyzes only pre-existing, publicly available images of public figures. This falls outside the U.S. Common Rule definition of human-subjects research (45 CFR 46.102). We do not redistribute the raw photographs. The public release (§7) comprises only derived per-identity 3D shape coefficients, the SHA-256-fingerprinted blind feature arrays, and analysis code. It does not include the source images themselves. Thus, the artifacts that enable reproduction do not republish identifiable facial photographs.

---

7. Data and code availability

You can find both pre-registrations on the Open Science Framework. They were timestamped prior to their respective confirmatory runs.

- **Registration 1 (n = 83):** OSF osf.io/smrxt - DOI 10.17605/OSF.IO/SMRXT. Frozen artifacts (SHA-256): estimator `confirmatory_analysis.py` (`578ec8f4…`, run unchanged via the `confirmatory_heldout.py` wrapper), held-out cohort `heldout_cohort_frozen.csv` (`ab664838…`), blind features `heldout_features.npz` (`da019428…`).

- **Registration 2 (n = 549, powered):** OSF osf.io/kj2rw - DOI 10.17605/OSF.IO/KJ2RW. Frozen artifacts (SHA-256): estimator `confirmatory_heldout.py` (`c22dcb076658bca3729c05554dc2c6dbe6` run byte-identical), held-out cohort `heldout_large_cohort_frozen.csv` (`e5c5786f…`), blind features `heldout_large_features.npz` (`2002e174…`).

The published SHA-256 fingerprints bind the estimator code, the frozen cohorts, and the blind feature arrays at registration time. This means any claim that the estimator, cohort, or features were altered after registration is falsifiable by re-hashing (see §2.3 for what this does and does not establish). The confirmatory analysis script, the per-coin result tables, the bootstrap and FDR outputs, the pose-adjustment outputs, and the power/sensitivity code are available alongside the registrations.

Competing Interests

The author developed, and holds a continuing interest in, the proprietary OPS-style typing system from which the trait-axis labels used here are derived. That system is the locus of the author's broader applied work and represents a potential commercial stake. Because this interest is material to a physiognomy-adjacent claim, we disclose it explicitly; we note that the result reported here is a **refutation** of the model's flagship prediction, a finding that runs against, not toward, any commercial interest in face-based typing. No other competing interests are declared.

Funding

This work received no external funding and was self-funded by the author. No funder had any role in the study design, analysis, interpretation, decision to publish, or preparation of the manuscript.

References

Companion papers: - Abdo M (2026a). Prenatal Hormonal Milieu as a Common Cause of Correlated Facial and Psychological Phenotype: A Multi-Axis Model and Its First Falsification Test. *PsyArXiv.* https://doi.org/10.31234/osf.io/akwrg - Abdo M (2026b). The Non-Shared Developmental Residual: What Twin Variance Implies for a Prenatal-Hormone Account of Correlated Phenotype. *PsyArXiv.* https://doi.org/10.31234/osf.io/qnmtg

Other references: - Agüera y Arcas B, Mitchell M, Todorov A (2018). Do algorithms reveal sexual orientation or just expose our stereotypes? *Medium.* - Guo J, Zhu X, Yang Y, Yang F, Lei Z, Li SZ (2020). Towards fast, accurate and stable 3D dense face alignment. *Proceedings of the European Conference on Computer Vision (ECCV)*, 152-168. doi:10.1007/978-3-030-58529-7_10. - Haselhuhn MP, Ormiston ME, Wong EM (2015). Men's facial width-to-height ratio predicts aggression: a meta-analysis. *PLOS ONE* 10(4):e0122637. doi:10.1371/journal.pone.0122637. - Kachur A, Osin E, Davydov D, Shutilov K, Novokshonov A (2020). Assessing the Big Five personality traits using real-life static facial images. *Scientific Reports* 10:8487. doi:10.1038/s41598-020-65358-6. - Kärkkäinen K, Joo J (2021). FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1548-1558. - Kosinski M (2017). Facial width-to-height ratio does not predict self-reported behavioral tendencies. *Psychological Science* 28(11):1675-1682. doi:10.1177/0956797617716929. - Kosinski M (2021). Facial recognition technology can expose political orientation from naturalistic facial images. *Scientific Reports* 11:100. doi:10.1038/s41598-020-79310-1. - Paysan P, Knothe R, Amberg B, Romdhani S, Vetter T (2009). A 3D face model for pose and illumination invariant face recognition. *Proceedings of the Sixth IEEE International Conference*

on Advanced Video and Signal Based Surveillance (AVSS), 296-301. doi:10.1109/AVSS.2009.58. - Radford A, Kim JW, Hallacy C, et al. (2021). Learning transferable visual models from natural language supervision (CLIP). *Proceedings of the 38th International Conference on Machine Learning (ICML)*, PMLR 139:8748-8763. arXiv:2103.00020. - Sorokowski P, Kowal M (2024). Relationship between the 2D:4D and prenatal testosterone, adult level testosterone, and testosterone change: meta-analysis of 54 studies. *American Journal of Biological Anthropology* 183(1):20-38. doi:10.1002/ajpa.24852. - Wang Y, Kosinski M (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology* 114(2):246-257. - Whitehouse AJO, Gilani SZ, Shafait F, et al. (2015). Prenatal testosterone exposure is related to sexually dimorphic facial morphology in adulthood. *Proceedings of the Royal Society B* 282:20151351. doi:10.1098/rspb.2015.1351.