A Forty-Year Audit of a Touchstone Study: A Critical Review of Shawn Carlson's "A Double-Blind Test of Astrology" (Nature, 1985)

Preface

Shawn Carlson's seven-page report in *Nature* 318 (5 December 1985), pp. 419–425, has been treated for four decades as the closest thing to a knockout punch in the empirical literature on astrology. It is cited in textbooks, encyclopaedias, science-museum exhibits, courtroom briefs, *Skeptical Inquirer* features, and the Wikipedia article on astrology. By Vidmar's estimate, the verdict that "astrology failed" has been transmitted, often verbatim, to "over an estimated 5 million people" Astrology-research astronlp (Vidmar 2008). For a paper of such cultural weight, the methodological scrutiny it has received from the mainstream skeptical community has been extraordinarily thin. The substantive criticisms have come almost entirely from psychometricians (notably H. J. Eysenck), retired research psychologists (Joseph Vidmar), and astrologers (Teresa Hamilton, Geoffrey Cornelius, Kenneth McRitchie, Robert Currey) — and from one outsider statistician, Suitbert Ertel of Göttingen, whose 2009 reanalysis published in the *Journal of Scientific Exploration* did not just chip at the edges of Carlson's design but reversed his principal numerical conclusion using his own data.

What follows is a methodologically neutral teardown of the paper. The point is *not* to argue that astrology works. It is to argue, as forcefully as the evidence permits, that *Carlson's paper does not show what it is universally said to show*; that several of its analytical decisions would, today, be classified as textbook examples of the practices that produced the replication crisis (post-hoc analysis switching, undisclosed sample-size changes, opportunistic significance thresholds, piecewise sub-sample analysis, and conclusion-drawing from null results); and that the published study, on its own data and on its own pre-specified protocols, supports a conclusion considerably more equivocal than the one Carlson, *Nature*, and the worldwide press disseminated.

---

1. The Study, Briefly

Carlson reports two "complementary" experiments (Carlson 1985, p. 419):

- **Part 1 (“self-selection”)** – test subjects were given three written natal-chart interpretations (their own plus two random alternatives drawn from the same Sun-sign group) and asked to rank them and rate them 1–10 for fit. A separate sub-test asked subjects to identify their own California Psychological Inventory (CPI) profile from three. InnerSelf Astrodivination
- **Part 2 (“astrologer matching”)** – up to 28 astrologers each received a natal chart together with three CPI profiles (one belonging to the chart’s owner, two random) and were asked to (a) rank them first/second/third best fit and (b) rate each on a 1–10 scale. Astrology News Service Astrologicalreviewletters

The CPI is an 18-scale personality inventory derived from 480 true/false items, Astrology News Service originally developed by Harrison Gough SciSpace (Carlson’s paper, p. 420). Carlson set significance at “2.5 standard deviations” above chance, “the level we had chosen to call ‘significant’” (Carlson 1985, p. 423; quoted in Ertel 2009, p. 135). The astrologers’ “predicted” minimum hit rate was 50 % (Carlson 1985, p. 425).

The study was funded and shepherded under the aegis of CSICOP — Carlson’s graduate advisor Richard Muller and the editor of *Nature* John Maddox were both CSICOP Fellows Correlation astrology-research (Currey 2011, pp. 8–9; Frazier 2009). Carlson, then a 21-year-old physics undergraduate, Astrology-research had an LBL preprint, LBL-20480, dated April 1983 (escholarship.org/uc/item/0b40b045); the paper was received by *Nature* on 11 April 1983 and accepted on 14 October 1985 astronlp +2 — an unusually long 30-month delay even by Nature’s standards (Vidmar 2008).

---

2. The Cardinal Methodological Sin: Post-Hoc Analytical Switching

The most surgical objection — and the one on which the rest of the critique pivots — concerns Carlson’s protocol for the astrologers’ matching task. On p. 425 of *Nature* he writes, in the only sentence in the paper that explicitly references his pre-data-analysis plan:

> *“Before the data had been analyzed, we had decided to test to see if the astrologers could select the correct CPI profile as either their first or second choice at a higher than expected rate.”* Journalofscientificexploration astrology-research (Carlson 1985, p. 425)

This is an admission, in print, that the pre-registered success criterion was **the combined first-plus-second-choice hit rate**. Under that criterion the astrologers ranked the correct CPI in 1st or 2nd place 86 times out of 116 — a hit rate of 0.741, against a chance expectation of 0.667. The binomial *Z* for that result is **1.61, p = 0.054 one-tailed**

journalofscientificexploration (Ertel 2009, pp. 129–130). By the conventional social-science threshold of $p < 0.05$ this is marginally significant; by Carlson's idiosyncratic 2.5-σ threshold it is not.

What did Carlson actually report? He published the first-, second-, and third-choice frequencies *separately* (his Table 2, p. 423), then dwelt on the third-choice frequency, observing that it was "consistent with chance" and concluding from that alone:

> *"Since the rate at which the astrologers chose the correct CPI as their third place choice was consistent with chance, we conclude that the astrologers were unable to chose [sic] the correct CPI as their first or second choices at a significant level."*
> journalofscientificexploration Astrology News Service (Carlson 1985, p. 425; quoted in Ertel 2009, p. 130)

This is a non sequitur. A null result on third-place hits does not imply a null result on combined-first-plus-second hits; in fact, with a forced-choice design, the two are mathematically linked, so a *high* combined first+second rate is *equivalent to* a *low* third-place rate. A first+second rate of 0.741 is not "consistent with chance" Astrology News Service — Ertel computed the binomial confidence intervals correctly (Newcombe/Wilson method) and showed the observed proportion sits at or beyond the upper 95 % limit journalofscientificexploration (Ertel 2009, p. 130). Carlson silently abandoned his own pre-registered analysis and substituted a piecemeal sub-analysis whose framing happened to support the null he set out to test.

This is a textbook instance of what Gelman and Loken (2013) call the "garden of forking paths": the protocol existed, the data existed, but the analytical *path* taken at the moment of publication was not the one specified ahead of time. In the modern replication-crisis vocabulary, the move would be called an undeclared deviation from a preregistered analysis plan (Gelman & Loken 2013). Carlson did not merely reframe; he ignored a sentence he himself had written ten lines above the conclusion. Ertel notes drily: "He ignored his own protocol Journalofscientificexploration without giving reasons" journalofscientificexploration Astrology News Service (Ertel 2009, p. 130).

---

3. The Rating Data: Discarded, Then Re-imported, Then Disaggregated

Carlson collected, by his own description, two distinct streams of data from each astrologer for each chart: (a) a rank ordering 1/2/3 of the three candidate CPIs, and (b) an *independent* 1–10 rating of how well each candidate CPI matched the chart. The astrologers' written instructions (preserved in unpublished materials examined by Ertel and Vidmar) ordered

them to perform the rating *first*, then the ranking Astrology News Service Astrologicalreviewletters — i.e., the rating was the higher-resolution, primary instrument and the ranking was a derived, coarser measure (Ertel 2009, p. 135 n. 7; Vidmar 2008).

Carlson's analytical handling of the rating data, which he displays in Figures 3 and 5, is again piecewise. Instead of pooling all astrologer ratings (308 ratings of 100 charts) and asking the obvious question — *do higher ratings predict authentic CPIs?* — he split the 1–10 ratings into three sub-samples according to which rank the rater had assigned in the *separate* ranking task, then fit three best-fit lines and observed that the slopes were "consistent with the scientific prediction of zero slope" journalofscientificexploration Scribd (Carlson 1985, p. 424). Astrology News Service This procedure makes no statistical sense. The two tasks were performed independently, used different sample sizes (116 charts for ranks, ~100 charts whose rating data could be reconstructed from his histograms), Astrologicalreviewletters and the rating scale carries information that the ranks cannot, Journalofscientificexploration since rank 1 of three nearly-identical CPIs and rank 1 of three obviously different CPIs convey different epistemic content.

When Ertel pooled the rating data correctly across the three sub-samples and computed a Kendall's tau correlation between the rating (1–10) and authenticity (1/0), he obtained:

> tau = 0.088, $Z = 1.78$, **p = 0.037** one-tailed, effect size ≈ 0.10 journalofscientificexploration (Ertel 2009, p. 131, Table 3).

In other words, when the variable Carlson originally instructed his astrologers to use *first* — and the variable that contains the most discriminating information — was analysed in the standard way, the astrologers' ratings of authentic CPIs were significantly higher than their ratings of non-authentic CPIs. The published paper's "no convincing evidence" conclusion is the artefact of an indefensible disaggregation. Currey's later multivariate regression, weighting the ratings by the frequency at each rating level, produces an even cleaner positive trendline (r = 0.57 across rating bins; Correlation Currey 2011, p. 16; Currey 2023).

There is, additionally, a discrepancy that Carlson himself never explained. His Table 2 reports 116/114/114 ranking decisions, with 40/46/28 hits; his Figures 3 and 5 plot ratings totalling 107/101/98 decisions, with 38/38/23 hits journalofscientificexploration (Ertel 2009, p. 135 n. 8). The shortfall — at least 9 fewer ratings than rankings in first place, 13 fewer in second, 16 fewer in third — implies that some astrologers completed only one of the two tasks. No mention of which, no mention of why, no robustness check.

There is a further, more troubling, discarded data set: the subject ratings. Carlson notes that he discarded the subjects' 1–10 ratings of their natal chart interpretations after seeing "the first few data envelopes," because a subject's first-choice section ratings tended to all be "first choice" too astrology-research (Carlson 1985, p. 422). But this was the only test of Part 1 that asked the *subjects* a meaningful continuous question, and discarding it after data

inspection — for reasons not fully spelled out — is exactly the kind of data-dependent sample-trimming that, if practised today, would prompt rejection or a demand for a robustness analysis. The raw data has never been made available so the decision cannot be audited astrology-research (see Currey 2011, p. 11).

---

4. The Significance Threshold: A Physics Convention Imported into Psychometrics

Carlson set his rejection threshold at "2.5 standard deviations" — equivalent to $p \approx 0.006$, journalofscientificexploration Astrologicalreviewletters two-tailed; $p \approx 0.012$ one-tailed. The conventional social-science threshold is $Z = 1.64$ / $p < 0.05$, Astrologicalreviewletters Astrology News Service a fact noted by Ertel (2009, p. 135) and remarked on by every subsequent commentator. Carlson's original LBL-20480 manuscript justified the choice by reference to physics conventions, but the relevant explanatory passage was *deleted by Nature in editing*. Per Vidmar:

> *"Since a positive astrological effect would be controversial, we decided at the outset to require a 2.5 standard-deviation increase over random choice to interpret the results as favoring the astrological hypothesis. (2.5 is a dividing line in physics experiments often used by skeptics before they are willing to accept a new or startling effect.)"* The portion in the parentheses, and the qualification "would be controversial", were removed in the *Nature* version astronlp (Vidmar 2008).

The asymmetry was not incidental: Carlson required 2.5σ for *both* sides — to reject the null *and* to reject astrology — but the empirical structure of the experiment (small N, low expected effect size, multiple sub-analyses) made the second prong much more easily reached than the first. With effect sizes typical of personality measurement (Cohen's *r* in the 0.10 range — comparable to the celebrated Gauquelin Mars effect, ES ≈ 0.03–0.07; journalofscientificexploration Ertel 2009, p. 133), the test power needed to clear $p \approx 0.006$ with N = 116 is essentially zero. By inflating the alpha threshold while leaving the sample small, Carlson designed an experiment that *could not pass* his own bar even if astrology produced an effect at the level Gauquelin's well-replicated planetary correlations claim. This is, in modern statistical terms, an underpowered confirmatory test masquerading as a refutation.

It is worth being precise about the Gauquelin comparison, which appears repeatedly in the rebuttal literature (e.g., Ertel 2009, p. 133; Currey 2011, p. 18; Ertel & Irving 1996). The Gauquelin/Mars-effect literature is itself contested; the point here is methodological, not substantive. Whichever side one believes, the sample sizes Gauquelin used were in the thousands precisely because the effects in question were small; journalofscientificexploration Carlson's design assumed effects an order of magnitude larger and required them to clear

an order-of-magnitude stricter alpha. If small effects exist in astrology, this design was guaranteed to miss them.

---

5. The Sample-Size Problem: Pre-Hoc 256, Post-Hoc 83/94/56/50/116/<28

In an excellent diagnostic question, Vidmar pursues the sequence of declared sample sizes:

> "We originally planned to be able to distinguish between the two hypotheses at a four standard deviation level.... Thus the total number of subjects was originally 256." astronlp (Carlson 1985, p. 421).

But:

- Part 1 self-selection: 83 test / 94 control astronlp (= 177 of an originally-intended 256, **a 31 % shortfall**) (Carlson 1985, p. 421);
- The CPI subject self-selection sub-test: 56 test / 50 control astronlp (a **59 % shortfall**) (Carlson 1985, Table 1, p. 422);
- Part 2 astrologer matching: **<28** astrologers, performing 116 ranking decisions journalofscientificexploration and 100 rating decisions (Carlson 1985, p. 421); the actual number of astrologers who completed assignments was never disclosed and is, by Vidmar's analysis of return rates, plausibly closer to 14–16 (Vidmar 2008). A surviving "FAVORING" letter from Carlson dated October 1981 — preserved by astrologer Erin Sullivan and reproduced in Vidmar 2008 — explicitly tells the participating astrologers that the experiment was at risk of failing "due to insufficient data returned." astronlp This is incompatible with the paper's stated 4σ-then-2.5σ design: at $Z = 4$ the required N is roughly four times Carlson's actual N.

Three separate things follow. First, *power was abandoned without disclosure*. Second, *the conclusion was upgraded — from "we couldn't show astrology" to "we showed astrology fails" — without a sample-size adjustment to match*. Third, "fewer astrologers than hoped for agreed to participate" astronlp (Carlson 1985, p. 421); some who agreed simply did not complete the task. Carlson himself acknowledges, indirectly, that the cohort of completers was smaller than advertised. The fact that *Nature* did not require him to state the actual completer count is one of the more striking peer-review failures of the paper.

---

6. Were the "Astrologers' Predictions" Real Predictions?

Carlson's strongest single rhetorical move is the claim that "the astrologers approved the design and predicted 50 percent as the 'minimum' effect they would expect to see" astronlp (Carlson 1985, p. 425). On this reading, astrologers picked a high bar, fell short, and convicted themselves out of their own mouths.

This is misleading on two grounds.

(a) Who actually predicted 50 %? Carlson does not tell us in the paper, but the unpublished materials make clear that the 50 % figure came from his three *advisory* astrologers — not from the participating astrologers, who were never collectively polled in advance, and not as the consensus of the National Council for Geocosmic Research (NCGR), which co-sponsored participant recruitment (Vidmar 2008; Hamilton 1986). Teresa Hamilton, one of the three named advisers and a CPI-trained M.A. psychologist who was thanked in the paper for "valuable suggestions," published in *Astropsychological Problems* 4(1), 1986, "Carlson followed none of my suggestions. I was never satisfied that the experiment was a fair test of astrology" Astrologicalreviewletters +2 (Hamilton 1986: 9). She also noted on first inspection of the materials that "the three profiles were often quite similar," Correlation astrology-research and considered the matching task "virtually impossible" astronlp (Hamilton 1986: 12). Hamilton resigned from the experiment. Correlation

(b) Was the prediction even a prediction? Astrologers (as Carlson's questionnaire records) routinely use intuition, context, multiple charts, and back-and-forth with the client. The "50 % minimum" was not an estimate derived from any quantitative model of natal-chart-to-CPI matching — it was a guess elicited under conditions where no one knew how the test would play out. Carlson's third hypothesis — that astrologers' confidence ratings would correlate with their accuracy — was another such elicited guess. None of these constitutes an *astrological* hypothesis in the sense that "Mars sextile Venus produces sociability" is an astrological hypothesis; they are confidence judgements, conflated by Carlson's writing with the substantive astrological prediction.

The asymmetry matters because failure to meet a strawman 50 % minimum is what Carlson's conclusion repeatedly leans on: "the astrologers' predictions proved to be wrong" astronlp (Carlson 1985, p. 425). If the prediction was not actually the astrologers' prediction, the rhetorical force of the sentence collapses.

---

7. The Self-Recognition Collapse and Its Implications for Part 1

The single most serious internal contradiction in the paper concerns the self-selection sub-test. Carlson's design hinged on a logical implication: if subjects can't recognise themselves, the experiment can't refute astrology. He writes:

> *"If subjects cannot recognize accurate descriptions of themselves at a significant level, then the experiment would show a null result no matter how well astrology worked."* Correlation (Carlson 1985, p. 424–425).

The result was that subjects could not, at the 2.5σ level, identify their own CPI profile from a set of three (test group: 25/56 first-choice authentic, 1.79σ above mean; chance 18.67) Astrodivination (Carlson 1985, p. 422; summarised in Cornelius 1986/2003). Carlson therefore *himself* concluded:

> *"We conclude that subject selection of astrologically derived information is a poor test of astrology."* Correlation astrology-research (Carlson 1985, p. 425).

This is a fatal admission for Part 1. Whatever Part 1 found about subjects' ability to identify their *own astrological interpretations* is uninterpretable, because Carlson cannot distinguish "astrology fails" from "self-recognition with this instrument fails." McGrew and McFall (1990) and Wyman and Vyse (2008) both built their later experiments around this very weakness — accepting that Carlson's Part 1 was inconclusive and that a non-CPI replacement was needed. Yet Carlson, having conceded Part 1 is uninterpretable, retained it in the paper, allowed its descriptive text to bolster the rhetorical case against astrology, and made conspicuous use of one peripheral Part 1 statistic — the 2.34σ excess of control-group "correct" interpretations — only to dismiss it as "statistical fluctuation." Cornelius noted, and Vidmar and Ertel agree, that this control-group anomaly would have been very hard to dismiss as fluctuation if it had happened in the test group and that the simpler explanation may be a clerical mix-up of the test and control envelopes (Cornelius 2003, Appendix 2; Vidmar 2008; Ertel 2009, pp. 131–132). Cornelius further observes that the control group ranked the (for them irrelevant) test-subject interpretation in *third* place only 18 of 94 times — a deviation Cornelius reports as 2.9σ, equivalent to $p < 1/520$ (Cornelius 2003, Appendix 2). This datum is in Carlson's published Table 1 but not commented on in his text. The point is methodological, not exegetical: the same standard of "fluctuation" can be applied or withheld at the analyst's discretion, and Carlson applies it precisely where it favours his preferred conclusion.

If Part 1 is genuinely uninterpretable (as Carlson himself says), then the paper's case rests entirely on Part 2 — which, on Ertel's defensible reanalysis using Carlson's own pre-registered criterion, *favours astrology at $p \approx 0.054$ in the ranking task and $p \approx 0.037$ in the rating task* (Ertel 2009, pp. 130–131). One can disagree with Ertel about the magnitude of the conclusion that follows, but it is no longer credible to claim that Part 2, properly analysed by Carlson's own protocol, refutes astrology.

---

8. The CPI Was the Wrong Instrument

Hans Eysenck — by H-index one of the three most-cited psychologists of the twentieth century, and a long-time experimentalist on personality measurement — published two reviews of Carlson's paper in early 1986 (in *Astropsychological Problems* 4(1) and *Correlation* 6(1)). Eysenck's verdict was uncompromising: *"the conclusion does not follow from the data"* (Eysenck 1986, summarised in Vidmar 2008).

Eysenck's specific objections to the use of the CPI are technical and damning:

- The CPI's 18 scales were not derived by factor analysis; they are "essentially arbitrary and subjective" because they were not chosen by ordinary psychometric techniques (Eysenck 1986, quoted in Ertel 2009, p. 128);
- The CPI manual itself warns that "it is important that scores on this test be interpreted by a competent psychologist" who has become "thoroughly familiar" with the instrument (CPI Manual; quoted in Vidmar 2008);
- The CPI scales generate **gender-differentiated profiles**, and Carlson withheld the sex of each CPI's owner from the astrologers to avoid identification leakage. This is a serious confound: it required astrologers to interpret an instrument *without one of its primary psychometric inputs* (Eysenck 1986: 8; Hamilton 1986: 10; Ertel 2009, p. 128);
- No participating astrologer was a licensed psychologist; the astrologers were given a 28-page "Interpreter's Syllabus" that contains, by Vidmar's count, only one example of a 2-scale combination out of 324 possible (Vidmar 2008). Subjects were given only a 2-page summary of what each scale meant — i.e., the astrologers and the subjects were both interpreting the CPI, but with different reference materials and different levels of preparation;
- McGrew and McFall (1990) themselves observed that "standard psychological tests, like the CPI, may not include the types of information that astrologers require to complete a matching task successfully" (McGrew & McFall 1990: 76).

The instrument issue compounds: if the CPI is a noisy mapping of the personality construct, *and* its gender component has been withheld, *and* the operators are not psychologists, *and* the subjects are 18–22-year-old college students whose CPI scores have an estimated 25-year test-retest median of $r = 0.58$ (CPP, cited in Currey 2011, p. 12 n. 14) — a population whose CPI profiles will be substantially overlapping for non-astrological reasons — then the experiment is a noisy measure of a noisy construct. The signal-to-noise ratio at the design stage made the chosen 2.5σ threshold all the more inappropriate.

Eysenck's broader observation in *Correlation* (1986) is harder to forget. He and Nias had reviewed dozens of empirical astrology studies for their 1982 book *Astrology: Science or Superstition?* and observed that, time and again, "the whole experiment had to be faulted because of quite elementary errors in the choice or interpretation of psychological

measuring instruments, errors which would be obvious to a first-year student of psychology." On Carlson he concluded: "Carlson selected a psychological test, which proved to be unsuitable, and which any competent psychologist could have predicted would prove so" (Eysenck 1986, summarised in Vidmar 2008). This is a serious indictment from a serious source, and the mainstream skeptical literature has not, in the four decades since, engaged it.

---

9. Three-Way Forced Choice and the Similar-Decoy Problem

Carlson's choice of a three-way (rather than two-way) forced choice, with the two decoy CPIs drawn at random from the pool of all subjects, has two related consequences.

First, **statistical power is lower for forced-choice three-way than for two-way**, the standard Thurstonian result (Thurstone 1927; Ertel 2009, p. 128). Vernon Clark's (1961) earlier studies of astrologer matching used pair comparison and reported significant effects; Carlson's design moved away from the format with the highest discrimination per trial.

Second, **the decoys are not just random — they are random within a homogeneous pool**. Because 70 % of subjects were Berkeley-area college students, with mean age ~ 28, English-speaking, predominantly recruited through campus advertising and the *Daily Californian* (Carlson 1985, p. 421; Vidmar 2008), the three CPI profiles offered to an astrologer were drawn from a population whose CPI profiles will be more similar than profiles from an age-and-occupation-diverse population. Hamilton's first-hand observation — "I was given some of these charts to match myself, and noticed immediately that the three profiles were often quite similar" — is not anecdotal whining; it is an observation about a *real* statistical property of the design (Hamilton 1986: 12). The Vernon Clark (1961) studies, by contrast, deliberately paired profiles known to be dissimilar; this is the standard fairness criterion for matching tests, and Carlson did not apply it.

The combined effect of (a) lower-power three-way format, (b) similar decoys, (c) a gender-blinded CPI, and (d) a homogeneous pool is to degrade the discriminability of authentic from non-authentic matches well below the threshold any astrological signal could plausibly clear at the chosen alpha — even before any consideration of whether astrology has anything to say about CPI scales in the first place.

---

10. Sample Composition: Belief Filter, Region, Age

Carlson screened out subjects who "strongly disbelieve" in astrology and those who had had a chart prepared (Carlson 1985, p. 421). The intent was to control demand bias on the

side of skeptical subjects. The unintended consequence is that the test-subject pool was biased toward neutrals and mild believers, who are *more* likely to find any of the three CPI profiles plausibly self-descriptive (the Forer/Barnum effect). For Part 1, this bias works *against* the astrological hypothesis (subjects would rate any of the three interpretations as plausible), reinforcing Carlson's own admission that Part 1 was uninterpretable.

The 70 %-college-student composition, with an overall mean age of about 28 (Carlson 1986; Currey 2011, p. 12), is more concerning. Most subjects had not yet experienced what astrologers call the "first Saturn return" (~ age 29–30) — the developmental milestone after which classical astrology claims personality traits become more fully expressed. From a purely psychometric standpoint, CPI test-retest correlations for adults over 25 years drop to a median $r \approx 0.58$ (CPP), suggesting a substantial portion of the variance in 18–22-year-old CPI scores is *not* stable trait variance. Whether one accepts the astrological framing or only the psychometric one, the pool was not a strong test of trait-level personality matching.

There were also asymmetric matching procedures: Carlson matched test/control pairs by sex, by age (to within at least three years), and by Sun sign — a clever procedural touch — but the matching was never independently audited, and the test-vs-control sample sizes (83 / 94 in Part 1, 56 / 50 in Part 2's CPI sub-test) are unequal in ways suggesting attrition was not random. Combined with the unaccounted-for gap between 116 ranking decisions and 100 rating decisions, the bookkeeping is not fully transparent.

---

11. The Curious Control-Group Anomaly

A telling secondary observation, due to Cornelius (2003) and amplified by Vidmar and Ertel: the control group in Part 1 — subjects given somebody else's astrological interpretation — *correctly* selected the "matched" partner's interpretation as their own first choice at 2.34σ above chance (Carlson 1985, p. 422). Carlson's interpretation was that this could not be an astrological effect (since the controls had not been given their own interpretations), could not be a Sun-sign effect (because the test group did not show it), and so must be "statistical fluctuation."

The probability of a 2.34σ deviation under a fair null is about 1 in 100; Cornelius further notes a separate 2.9σ deviation in third-place rankings for control subjects, $p < 1/520$ (Cornelius 2003, Appendix 2). Cornelius offers a "context-psi" explanation; Vidmar offers the more parsimonious hypothesis that Carlson's clerical chain inadvertently transposed test and control envelopes, since the matching procedure (by sex/age/Sun-sign) was complex and the data envelopes were re-coded multiple times (Vidmar 2008; Ertel 2009, p. 132). Both are speculative. The point is methodological: the same paper that treats a 1.61σ deviation in Part 2 as "consistent with chance" treats a 2.34σ deviation in Part 1 as "statistical fluctuation" — i.e., the analyst's choice of which deviations are real and which are

noise is doing the work.

---

12. Publication-Process Concerns

Several procedural irregularities deserve mention not because they invalidate the data but because they are documented in the public record and bear on the study's evidentiary weight.

- **Peer review.** Carlson's paper appeared in *Nature*'s "Commentary" section, the section in which articles are not subject to standard peer review and are placed at the editor's discretion (Vidmar 2008). Carlson has stated the paper "survived a rigorous peer review that included a famous psychologist whom I will reveal in a later publication" (Carlson 2009, in Skeptico); the famous psychologist has not, to my knowledge, been revealed. The conventional 5-page Commentary limit was extended for this 7-page paper.
- **Editorial deletions.** *Nature*'s editing removed the sentence justifying the unusual 2.5σ threshold and the sentence noting that "a positive astrological effect would be controversial" (Vidmar 2008) — i.e., precisely the disclosures that would have flagged the threshold's unusual nature for readers.
- **Editorial alignment.** *Nature*'s editor John Maddox was a CSICOP Fellow (Frazier, *Skeptical Inquirer* 2009), as was Carlson's funding mentor Richard Muller. CSICOP's founder Paul Kurtz "encouraged" the project (Kurtz 2006). This is not, by itself, evidence of misconduct, but it is a relevant disclosure for an experiment whose conclusion was treated as definitive partly because it was published in *Nature*.
- **Submission-to-publication delay.** 30 months elapsed between submission and publication (escholarship.org/uc/item/0b40b045 for the LBL preprint date; *Nature* received 11 April 1983, accepted 14 October 1985). *Nature*'s author guidelines (then and now) specify that authors "are usually informed within a week if the paper is not being considered" (Vidmar 2008). The delay coincides with the post-1981 "STARBABY" / Mars-effect crisis at CSICOP (Rawlins 1981; Currey 2011, pp. 9–10), in which CSICOP's attempts to refute Gauquelin's planetary correlations had failed. None of this constitutes evidence of misconduct, but it is an important contextual feature for a paper that made dispositive claims about a research field its sponsors had a documented institutional interest in discrediting.
- **Carlson's own disclosure during data collection.** A signed but undated October 1981 letter from Carlson to the participating astrologer Erin Sullivan — preserved by Sullivan and reproduced in Vidmar 2008 — states: *"We have completed much of the data analysis on the data which has been returned to us. We are very near interpreting the*

results as FAVORING the astrological thesis. Near, but not there yet." This is incompatible with the published claim (Carlson 1985, p. 422) that "at no time during the data collection did the experimenter have access to any information relating subjects' identities to code numbers. This control was abandoned only when all the data had been collected and the methods of analysis had been established." Either the letter exaggerated, or the published statement does. Unblinded interim analyses of the kind described in the letter, if performed, would substantially weaken the double-blind claim (Vidmar 2008).

---

13. The Reanalyses: Ertel (2009), Currey (2011), and Their Limits

The two principal reanalyses are owed to:

- **Suitbert Ertel** ("Appraisal of Shawn Carlson's Renowned Astrology Tests"), *Journal of Scientific Exploration* 23(2) (2009): 125–137. Ertel was emeritus professor of psychology at Göttingen, an established methodological critic of researchers on *both* sides of the astrology debate (including the Gauquelins themselves, whom he criticized for sampling artefacts; Ertel 1988). His credentials in this domain are the strongest of any critic.
- **Robert Currey** ("U-turn in Carlson's astrology test"), *Correlation* 27(2), July 2011, pp. 7–33. Currey is a practising astrologer; his contribution adds a multivariate regression on the rating data and synthesises the Vidmar/Ertel critiques.

Ertel's results, summarised:

Test	Carlson's analysis	Ertel's analysis
Part 2 ranking (1st+2nd hits)	"consistent with chance"	86/116 hits, $Z = 1.61$, **p = 0.054**
Part 2 ratings (1–10)	"consistent with zero slope"	tau = 0.088, $Z = 1.78$, **p = 0.037**, ES ≈ 0.10

Ertel 2009, pp. 129–131.

Ertel's reanalysis is methodologically more defensible than Carlson's for three reasons. First, it uses the *pre-specified* analytical rule (combined first+second choice) that Carlson himself stated. Second, it pools the rating data across the three rank sub-samples — which is what should always be done with two independent measurements of the same underlying construct. Third, it uses a non-parametric correlation (Kendall's tau) appropriate to ordinal

rating data. None of these is a sleight of hand; all are textbook moves.

That said, Ertel's reanalysis is not a free pass:

- **The p-values are marginal.** $p = 0.054$ is not below the 0.05 threshold; $p = 0.037$ only modestly is. Ertel himself is admirably cautious: "The results are regarded as insufficient to deem astrology as empirically verified, but they are sufficient to regard Carlson's negative verdict on astrology as untenable" (Ertel 2009, abstract, p. 125). The conclusion is asymmetric — the data refute Carlson's strong claim, not the null.
- **Multiple-comparisons concerns work both ways.** Ertel chose the analyses that (in his judgement) Carlson should have done; given the freedom that the original disorganised paper allows, an alternative reanalyst could choose other analyses. The "garden of forking paths" applies to reanalyses too. That said, Ertel's chosen analyses are pre-specified by Carlson's own protocol document.
- **Effect sizes are very small.** ES ≈ 0.10 is comparable to the smallest effects in the social sciences and to the Gauquelin Mars-effect range. These are not effects you can reliably detect in N = 116. So the picture, even on Ertel's analysis, is consistent with: (a) astrology produces a small effect; (b) astrology produces no effect, but Carlson's data have a small artefact in the favourable direction; or (c) something between. The data alone cannot adjudicate.
- **Currey's regression on rating data**, weighted by frequency at each rating bin, produces $r = 0.57$ across rating bins (Currey 2023). But this is a between-bin correlation, not a per-trial effect size, and conflating the two is easy. Ertel's tau (0.088) is the more honest figure at the level of individual trials; Currey's r describes the consistency of the rating-vs-hit-rate trend across the rating scale, which is impressive within the data but does not by itself demonstrate a large per-trial effect.

The cleanest summary of the post-Ertel state of the data is therefore: *Carlson's published conclusion is not supported by his own data; but the data also do not constitute strong evidence of an astrological effect.* Both directions of overclaim should be resisted.

Dean and Kelly's response in *Understanding Astrology* (2022, p. 574) acknowledges the positive trendline and an effect size of $r = 0.385$ across rating bins but argues $p = 0.27$ (Dean et al. 2022, summarised in Currey 2023; see also Currey 2024 review). This is a different statistical question: Dean treats each rating-bin point as one observation and ignores the frequency weights, which Currey argues is the wrong approach. The dispute is unresolved and remains the live frontier of the methodological debate.

---

14. Other Critiques Worth Naming

- **Joseph Vidmar**, "A Comprehensive Review of the Carlson Astrology Experiments," *Correlation* 26(1), 2008, pp. 14–32 (Vidmar 2008 full text). A retired Louisiana professor of psychology with editorial experience in psychology journals; he located four surviving participating astrologers, including Erin Sullivan, who had retained the original instructional materials in storage. Vidmar's contribution is largely archival and procedural rather than statistical; he documents the unblinded interim analysis in Carlson's October 1981 letter, the discrepancies between the LBL preprint and the published paper, the confidentiality issues around the CPI's restricted-use copyright, and the misattribution of the 50 % "minimum" prediction. Vidmar's article is partisan in tone and includes a speculative comparison of the experimental design to the three-card-monte sleight-of-hand confidence game (a vocation Carlson reportedly worked at to support himself in adolescence; Carlson, in Boston Globe, 2005). The comparison is rhetorical; the documentary evidence Vidmar cites is not.

- **Geoffrey Cornelius**, "The NCGR-Berkeley Double-Blind Test of Astrology," *Astrology Quarterly* 59(4), 1985/86, and Appendix 2 of *The Moment of Astrology* (rev. ed. 2003). Cornelius noted the control-group anomalies that Vidmar and Ertel later took up (Cornelius 2003, Appendix 2).

- **Teresa Hamilton**, "Critique of the Carlson Study," *Astropsychological Problems* 4(1), 1986, pp. 9–12 (reprinted in *Kosmos*, Spring 1986). The most authoritative inside critique, by a participating astrologer and CPI-trained M.A. psychologist. Hamilton documents that her methodological recommendations were ignored.

- **John H. McGrew and Richard M. McFall**, "A Scientific Inquiry into the Validity of Astrology," *JSE* 4(1), 1990, pp. 75–83 (full text). A skeptic-aligned attempted replication that explicitly took up Carlson's self-recognition problem and replaced the CPI with case files, photos, the Strong-Campbell, and 16PF. McGrew and McFall reported chance-level matching, but their design imposed a 23-way matching task on six astrologers from a small Indiana group with no claimed accreditation, used subjects all aged 30–31 (to avoid age-leakage from photographs), and used birth-data clustered around the 1958 Saturn-Neptune conjunction — i.e. the homogeneity problem Carlson had was made *worse* (McRitchie 2014). The McGrew-McFall paper does not, on careful reading, vindicate Carlson; it concedes the inadequacy of Part 1 and substitutes an arguably worse design.

- **Wyman & Vyse**, "Science Versus the Stars," *Journal of General Psychology* 135(3), 2008, pp. 287–300 (PubMed 18649494). A near-replication using the NEO-FFI and computer-generated chart text. With N = 52, the study found subjects could identify their NEO-FFI profile but not their astrological summary. McRitchie (2014) catalogues several procedural problems including a non-blinded "knowledge and beliefs" questionnaire and use of computer-generated rather than astrologer-generated profiles

(McRitchie 2014b). A 2018 attempted re-replication did not reach its preregistered N=50 (Wagenmakers et al. discussed in ResearchGate).

- **Garry Phillipson** has written on the methodological–philosophical status of the Carlson study from a perspective that takes astrology-as-divination rather than astrology-as-natural-science as the analytic frame; Phillipson argues that Carlson's design assumes a foundationalist/realist epistemology that astrologers themselves often reject (e.g., Phillipson 2006, *Correlation* 23(2); Phillipson 2020 PhD thesis). Phillipson's critique is largely orthogonal to the statistical one, but it usefully reframes the question of what a "fair test" of astrology would even consist of.
- **Kenneth Irving** (longtime co-author with Ertel on Gauquelin matters; *The Tenacious Mars Effect*, 1996) has not produced a dedicated published critique of the Carlson study, but his methodological position is reflected in The Astrology Podcast Ep. 173 (2018) and in the institutional history of CSICOP-era research practices around astrology.
- **Kenneth McRitchie** has written a series of articles in *ISAR International Astrologer*, the *Journal of Consciousness Studies* (2016), and the *Journal of Scientific Exploration* (2022, 2023) developing the implication of Ertel's reanalysis and engaging Dean and Kelly directly (McRitchie 2016; McRitchie 2023 in JSE 37(3)). McRitchie is a partisan voice; his statistical work is largely synthesis rather than independent reanalysis.

---

15. The Mainstream Skeptical Response to Ertel: A Brief Audit

A reasonable expectation, given the gravity of Ertel's claim — that a flagship CSICOP-aligned *Nature* paper does not, on reanalysis using its own pre-specified protocol, support the conclusion drawn — would be that *Skeptical Inquirer*, the *James Randi Educational Foundation*, the Committee for Skeptical Inquiry, or the broader scientific media engage the reanalysis substantively. As of the present writing, that engagement has been minimal. The principal published rebuttals to Ertel come from Geoffrey Dean and Ivan Kelly, mostly in the four editions of their massive critical compendium (*Astrology under Scrutiny*, 2013; *Tests of Astrology*, 2016; *Understanding Astrology*, 2022) and in *JSE* exchanges (Dean and Kelly 2023, McRitchie 2023). Dean's central counter is that the rating-vs-hit-rate trendline, when computed without weighting, yields $p = 0.27$ (Dean et al. 2022 p. 574; reviewed in Currey 2023). The blogosphere reaction has been, charitably, thin: the only substantive public skeptical engagement traceable through 2009–2025 is on the "Skeptico" blog (Skeptico 2009, comment thread), and Carlson himself made a brief remark there to the effect that Ertel had performed a post-hoc analysis ("A scientist never figures out how to analyze his data after the fact with all the data in view"; quoted in Currey 2011, p. 19). The remark is doubly ironic: Ertel applied the analysis Carlson himself had pre-specified.

In short: a *Nature*-published study from a CSICOP-aligned project, whose conclusion has been reanalysed by a Göttingen statistician using the original protocol and shown not to support the published verdict, has not — over fifteen years — produced a published reanalysis from CSI/CSICOP methodologists, nor a Carlson-authored response in a peer-reviewed venue. *That* is a sociological feature of the scientific literature on astrology worth noting in its own right.

---

16. Sub-Question Summaries

To map the user's specific sub-questions onto the analysis above:

(1) Did Carlson actually pre-register his analysis plan, and did he deviate from it? Yes and yes. He pre-specified, in print on p. 425, that the success criterion was the combined first+second-choice hit rate. He then analysed, instead, third-choice hit rates and inferred from a null result there to a null result on first+second. The unpublished instruction sheet to astrologers also directed them to perform the rating before the ranking; Carlson's published analysis treats the rating as a coarser version of the ranking, not as an independent measurement. Both deviations are textbook examples of the practices the modern preregistration movement was designed to prevent (Ertel 2009, pp. 129–131; Currey 2011, pp. 13–17).

(2) What were the actual numerical results when analysed by the originally-planned method (rating scales) vs. the actually-used method (ranking)?

- As Carlson analysed (third-choice frequency, sub-sample analyses, 2.5σ threshold): null on all measures.
- On Carlson's *pre-specified* rule (first+second choice combined): 86/116, $p = 0.054$.
- On the rating data pooled across rank sub-samples: tau = 0.088, $p = 0.037$, ES ≈ 0.10 (Ertel 2009, pp. 129–131).

(3) Is Ertel's reanalysis methodologically sound, or does it have its own problems? Sound on the central question (apply the pre-specified rule). Some technical caveats: the rating-data reconstruction depends on plot-reading from Carlson's Figures 3 and 5, and there are unexplained 9–16-decision shortfalls between the ranking and rating totals; tau is appropriate but the small sample makes the *p*-values fragile; Ertel's selection of these particular analyses has its own forking-paths character, though it is constrained by Carlson's own protocol. Ertel himself draws a deliberately limited conclusion: not that astrology is empirically verified, but that Carlson's negative verdict is "untenable" (Ertel 2009, p. 134).

(4) What is the appropriate Bayesian or frequentist conclusion given all the data? A frequentist analyst applying Carlson's preregistered protocol obtains $p \approx 0.04$–0.06 in favour of the alternative. A Bayesian with a heavily skeptical prior on astrology will update toward astrology working but not nearly enough to shift posterior odds; with a flat prior, the data move modestly toward the alternative. Either way, the appropriate conclusion is *not* that the experiment refutes astrology. It is that the experiment is too small and too poorly instrumented to discriminate "no effect" from "small effect" — and that, on the data Carlson collected and the protocol he wrote, the small-effect hypothesis is not less consistent with his data than the null. The failure mode of the paper is overclaim, not error.

(5) Did the paper accurately represent the predictions of the participating astrologers vs. the advisory astrologers? No. The 50 % "minimum" prediction was elicited from the three advisory astrologers, not from the 28 participating astrologers, and one of the three (Hamilton) publicly disavowed the experiment. Two of the three (per Vidmar's reconstruction) protested after publication; the third was terminally ill at the time of publication (Vidmar 2008; Hamilton 1986).

(6) What is the current scientific consensus on the Carlson study?

- *Among mainstream skeptics:* The 1985 verdict still stands. Citation in textbooks, popular books, and Wikipedia treats the paper as a definitive negative result. The Ertel (2009) and Currey (2011) reanalyses are not engaged in mainstream skeptical literature; CSI has not issued a formal response.
- *Among parapsychologists/astrology researchers:* Ertel's reanalysis is treated as decisive; Carlson's verdict is regarded as overturned. This is a partisan position but its statistical core is not in serious dispute among those who have engaged the data.
- *Among statisticians who have looked at it:* The handful (Ertel; the *JSE* peer reviewers; Geoffrey Dean, who is by training an analytical chemist with a strong statistics background; the various authors collected in Dean et al. 2022) who have done independent reanalyses largely agree that Carlson's analysis is non-standard and that his conclusion outruns his data. They disagree on whether Ertel's *replacement* analysis is fully convincing. There is no third-party reanalysis from a non-aligned statistician in a mainstream statistics journal.

---

17. Final Assessment

The Carlson paper is best read as a study about the dangers of analytical drift — how a not-unreasonable initial design can be undone by a sequence of small, individually defensible-looking decisions that compound in one direction. Specifically:

1. Pre-specified first+second-choice analysis was abandoned in favour of a third-choice-only analysis.
2. Pre-specified rating data, instructed to be collected first, was treated as a derivative of the ranking and disaggregated into three under-powered sub-samples.
3. Sample size was allowed to drop to less than half the originally planned 256 without a compensating adjustment to the alpha threshold.
4. The alpha threshold was set at 2.5σ — a physics convention, justified by a sentence the journal then deleted — that is essentially impossible to clear given small effect sizes and a reduced N.
5. Part 1 was retained for rhetorical purposes despite Carlson's explicit acknowledgement that its self-recognition collapse rendered it uninterpretable.
6. The CPI was used by non-psychologists in an unconventional matching task, with the sex of the test-takers withheld — invalidating, by the CPI manual's own warning, the instrument as a measure.
7. The 50 % "minimum prediction" attributed to the participating astrologers came from three advisers, one of whom resigned in protest.
8. Procedural transparency (raw data, exact astrologer count, ranking-vs-rating discrepancies, rationale for dropping subject ratings) is below the standard a 2025 *Nature* paper would face.

Independently, when analysed using Carlson's *own* pre-specified protocol, the data point in the *opposite direction*: marginal evidence in favour of an astrologer-matching effect at the $p = 0.04–0.06$ level, with a small effect size ($r \approx 0.10$).

The honest conclusion is the one Ertel himself drew: *the data are insufficient to verify astrology, but they are sufficient to make Carlson's negative verdict untenable*. (Ertel 2009, p. 134)

It is striking that a paper with these properties has been treated for forty years as the definitive empirical refutation of a field of inquiry. Whatever one believes about astrology — and the existence of a small effect at $r \approx 0.10$ in Carlson's data is a long way from the "fundamental thesis of natal astrology" the paper ostensibly tested — the paper does not say what its citing literature says it says. This is, in the end, a story about the methodological frailty of a culturally weighty study; about the importance of preregistration, of distinguishing confirmatory from exploratory analysis, of recording and adhering to success criteria; and about the slowness with which a flagship scientific verdict, once issued, is corrected on the basis of statistical reanalysis even when the reanalysis comes from a credentialled source and uses the original protocol.

The Carlson paper does not show that astrologers cannot match natal charts to CPI profiles. It shows that, on a small sample, with a non-standard test, a non-standard analytical pipeline, and a non-standard significance threshold, an under-powered experiment failed to clear an artificially high bar. To call that "a strong case against natal astrology as practised by reputable astrologers" is to confuse the *failure to find* with the *finding of failure* — the most basic null-hypothesis fallacy in introductory statistics, and the one Carlson's own paper, in its concluding sentences, commits.

Bibliography of cited primary and secondary sources

- Carlson, Shawn. (1985). "A double-blind test of astrology." *Nature* 318: 419–425. LBL-20480 preprint, April 1983; PDF via R. Muller, LBL.
- Cornelius, Geoffrey. (1986). "The NCGR-Berkeley Double-Blind Test of Astrology." *Astrology Quarterly* 59(4): 183–8. Expanded as Appendix 2 in *The Moment of Astrology* (rev. ed., Wessex Astrologer, 2003). Summary at Astrodivination.
- Currey, Robert. (2011). "U-turn in Carlson's Astrology Test?" *Correlation* 27(2): 7–33. Full PDF.
- Currey, Robert. (2023). "The Carlson Test continues to support Astrology." *Correlation* 35(2): 76–83. Full text.
- Dean, Geoffrey, Mather, Arthur, Nias, David, & Smit, Rudolf. (2022). *Understanding Astrology: A Critical Review of a Thousand Empirical Studies 1900–2020*. AinO Publications, Amsterdam. Discussed in Currey 2024 review.
- Dean, Geoffrey & Kelly, Ivan. (2003). "Is Astrology Relevant to Consciousness and Psi?" *Journal of Consciousness Studies* 10(6–7): 175–198.
- Ertel, Suitbert. (2009). "Appraisal of Shawn Carlson's Renowned Astrology Tests." *Journal of Scientific Exploration* 23(2): 125–137. Full PDF.
- Ertel, Suitbert & Irving, Kenneth. (1996). *The Tenacious Mars Effect*. Urania Trust, London.
- Eysenck, Hans J. (1986). "Critique of 'A double-blind test of astrology'." *Astropsychological Problems* 4(1): 7–8 and *Correlation* 6(1): 15–16. Extracts in Vidmar 2008.
- Gelman, Andrew & Loken, Eric. (2013). "The garden of forking paths." Columbia working paper.
- Hamilton, Teresa W. (1986). "Critique of the Carlson Study." *Astropsychological*

Problems 4(1): 9–12; reprinted in *Kosmos*, Spring 1986: 26–31.

- McGrew, John H. & McFall, Richard M. (1990). "A Scientific Inquiry into the Validity of Astrology." *Journal of Scientific Exploration* 4(1): 75–83. Full PDF.
- McRitchie, Kenneth. (2011). "Support for Astrology from the Carlson Double-Blind Experiment." *ISAR International Astrologer* 40(2): 34–39. Reprint.
- McRitchie, Kenneth. (2016). "Clearing the Logjam in Astrological Research." *Journal of Consciousness Studies* 23(9–10): 153–179. PhilArchive.
- McRitchie, Kenneth. (2023). "Defense of Emergent Effects in Astrology Research: Rebuttal of Dean and Kelly (2023)." *Journal of Scientific Exploration* 37(3): 576–579.
- Phillipson, Garry. (2000). *Astrology in the Year Zero*. Flare Publications, London.
- Phillipson, Garry. (2020). *Astrology and Truth: A Context in Contemporary Epistemology*. PhD thesis, University of Wales Trinity Saint David. Repository.
- Rawlins, Dennis. (1981). "sTARBABY." *FATE* 34. Online.
- Vidmar, Joseph E. (2008). "A Comprehensive Review of the Carlson Astrology Experiments." *Correlation* 26(1): 14–32. Online text.
- Wyman, Alyssa Jayne & Vyse, Stuart. (2008). "Science Versus the Stars: A Double-Blind Test of the Validity of the NEO Five-Factor Inventory and Computer-Generated Astrological Natal Charts." *Journal of General Psychology* 135(3): 287–300. PubMed.