

The Sign of Cannibalization Feedback

Does cannibalization intensity rise with supply-driven unavailability? Three panels of published fleet statistics, nine specifications, one sign and no magnitude

Dayan Ibrar

THREE ESTIMATES FROM PUBLISHED OPERATIONAL STATISTICS, PLUS A FOURTH REPORT USED ONLY TO CHECK THEM · PREPRINT, NOT PEER REVIEWED

Three estimates from published operational statistics, plus a fourth report used only to check them. We ran no experiment on any fleet, and no number here is simulated. Every value below is one of three things, and the text says which: a quantity transcribed from a published table and checked against that table, a quantity digitised from the vector geometry of a published chart that prints no values, or a statistic computed from those by a stated estimator. No calibration of any model is offered.

Abstract

An operator deciding where to spend the next unit of maintenance capacity may want to know one thing that needs no model: whether cannibalization intensity rises or falls as more of a fleet goes down for want of parts. This paper estimates the sign of that relation from published operational statistics, on three panels that share no pair of variables. On GAO-01-771's Navy tables — eight fiscal years, two aircraft types, 16 paired observations — the within-type slope of cannibalizations per 100 flying hours on supply-driven unavailability is $+1.089$ ($t = +4.79$). That significance does not survive giving each type its own linear trend, because 83.3 per cent of what identifies it is coincident trend, so the estimate is replicated twice. On 38 Air Combat Command category-years the same attack leaves $+0.813$ ($t = +4.49$), and on 138 monthly airlift observations digitised from GAO-01-495R it leaves $+0.500$ ($t = +3.59$). The sign is positive in all nine specifications the three panels admit, six of them clearing five per cent under independent errors. No magnitude is established: no two panels measure the same pair of quantities, so none may be pooled, none can calibrate a model, and none identifies a direction of causation. No panel has enough clusters for cluster-robust inference, and every t here assumes independent errors.

1. Introduction

1.1 The question

A maintenance organization running a fleet of aircraft can observe two things about itself without any model at all: how much of the fleet is down for want of a part, and how hard its maintainers are cannibalizing — taking a part off one airframe to return another to service. The practice is documented across the services, and the GAO has reported that it is both widespread

and inconsistently tracked, with the Army at that time not recording it service-wide at all [1].

The question here is which way those two quantities move together within one fleet. If cannibalization intensity falls as supply-driven unavailability rises, the practice is self-limiting and a shortfall stays roughly where it lands. If it rises, a shortfall creates demand, and the fleet carries positive feedback on its own disabled fraction. The two cases differ in what an operator should expect from a unit of additional supply, and the distinction is prior to any particular model of the feedback.

A theoretical motivation for the question does exist, and it is developed elsewhere. Everything structural in *A Curvature Criterion for Cannibalization Feedback* reduces to one sign: its criteria all compare $2b'(x) + x b''(x)$ against a threshold, where $b(x)$ is the share of disabled assets disabled by someone else's shortage, and for the linear law $\beta(u) = \beta_0(1 + \kappa(1 - u))$ that quantity is $2\beta_0\kappa$, so whether readiness is concave — whether greedy keeps its guarantee — turns on whether κ is positive and large enough. That note assumes it is positive. Our own discrete-event simulator, reported in *Readiness at the Critical Point*, measured it negative. Neither settles the question, because the first is an assumption and the second is a property of a donor policy we wrote.

The sign is an empirical quantity, and this paper estimates it from published operational statistics. The estimand is a within-unit slope of a published cannibalization rate on a published unavailability rate, and what is reported is the sign of that slope and its significance. A reader who rejects the model can still read the estimate, and a reader who accepts the model gets from it only a sign.

1.2 What this paper establishes

The propensity rises with the disabled fraction on real fleets (estimated from operational data). Within unit, cannibalization intensity rises with supply-driven unavailability on three independent panels and in all nine specifications they admit, and it survives removing unit-specific linear trends on two of the three — $t = +4.49$ on 38 Air Force category-years and $t = +3.59$ on 138 airlift aircraft-months, where the Navy panel's $t = +4.79$ does not survive it. Those t statistics assume independent errors, which is the only basis on which all three panels are comparable and is not defensible on all three: the airlift figure falls to $+1.90$

under a year of lags, and no panel here has enough clusters for cluster-robust inference. The reports' own documented underreporting biases every one of those estimates toward zero. This is the sign the model of *A Curvature Criterion for Cannibalization Feedback* assumes, and it is the only evidence we have bearing on that assumption from outside the model.

On real fleets the sign is positive, across two services, two decades and three panels. GAO-01-771 reports, for the same Navy fleets over eight fiscal years, both the share of aircraft down for want of a part and the rate at which maintainers cannibalized; the within-type slope of the second on the first is $+1.089$, $t = +4.79$. That significance does not survive giving each type its own linear trend, because 83 per cent of what identifies it is coincident trend, so we replicate twice. On Air Combat Command's independent tables for 1980 to 1999 the same attack leaves $+0.813$ at $t = +4.49$, with 11 per cent of the identifying covariance trend. On the monthly charts of GAO-01-495R — 138 paired observations for the C-5, KC-135 and KC-10, eight times the first panel's size and carrying the supply-driven regressor itself — it leaves $+0.500$ at $t = +3.59$, with 25 per cent trend — though that panel is the only one of the three long enough to measure serial correlation, and correcting for it leaves that figure marginal ($+2.18$ at four months of lags, $+1.90$ at twelve). **The sign is positive in all nine specifications the three panels admit, and survives detrending in two of the three.** Each panel is weak in a different place and the places do not overlap: the Navy magnitude is not identified, the Air Force regressor is total rather than supply-driven unavailability and its primary document is one we have not read, and the airlift values were digitised from chart geometry because the charts print no numbers. No two share both variables, so none may be pooled, none can calibrate the model, and none identifies a direction of causation. What the three establish jointly is a sign, which is what the criteria need. A fourth GAO report, which pairs nothing and is used only to check them, agrees exactly with the first on the one cell both print, puts the digitised values within five and eleven per cent of an independently printed table, and finds no association between average fleet age and cannibalization — which weighs against the obvious alternative explanation of the sign without excluding it, since those cross-sections are small enough that an effect of up to about 1.5 standard deviations across the age range remains inside the interval.

2. Related work

Cannibalization has been modelled for forty years, and the modelling literature does not agree on the sign. Fisher and Brennan simulated competing cannibalization policies under spares, repair and resource constraints and found the question to be a policy comparison rather than a fixed hazard, with the outcome turning on how long a part removal takes relative to a repair [3]. Gaver, Isaacson and Abell built designated cannibalization into Air Force recoverable spares requirements and concluded that it is cost-effective, permitting lower safety stock at the

same availability [4]. Sheng and Prescott model fleet maintenance as a hierarchical coloured Petri net, compare no cannibalization against unrestricted cannibalization and against a designated donor airframe, and optimize the number of donor airframes and how long each serves [5], later extending that model to spare inventory and preventive maintenance [6]. What this literature establishes is that the practice has a benefit side and that its effect is a function of policy.

The nearest prior work runs the relation the other way. Trost, Ye, Hoover, Goodwyn and Jondrow give the Navy a probabilistic framework for why cannibalizations occur and what drives their rate, and use it to predict the cannibalization rate needed to reach a specified readiness goal given expected customer wait time for spare parts [12]. That is the inverse of the question here: they ask what rate a readiness target implies, where this paper asks which way an observed rate moves with observed supply-driven unavailability. We cite them for that framing, which is as much as their abstract supports and as much as we have read; we have not checked their numbers and make no claim about them. With that said, what the modelling literature does not supply is an estimate of the *sign* of the within-unit relation taken from a fleet's own published statistics, and that is what this paper is.

The inventory literature does not supply that estimate either, because it does not need it: Sherbrooke's METRIC model and most of the multi-echelon work following it treat demand as exogenous Poisson [2], and the quantity at issue here is the part of demand that is endogenous. The nearest prior treatment of feedback of this shape is outside sustainment altogether, in the power-systems line of Dobson, Carreras, Lynch and Newman, who model cascading blackouts as a branching process whose offspring mean rises with system loading, and who treat criticality and finite-size saturation carefully and against measured data [7]. The mechanism at issue here is the same shape as theirs and we claim no novelty for it; what this paper adds is a sign for it, estimated on fleets rather than assumed.

3. Three public series reporting both sides of the law

The model needs two things measured on the same fleet over the same period: how much of it is down for want of a part, and how hard maintainers are cannibalizing. GAO-01-771 [8] reports both for Navy aircraft over fiscal years 1993 through 2000, sourced to the Deputy Chief of Naval Operations for Fleet Readiness and Logistics. Its Table 6 gives the share of aircraft not mission capable due to a supply shortage, by type. Its Table 7 gives cannibalizations per 100 flying hours, for the Navy in aggregate and for the EA-6B and the F-14.

We transcribed those tables and the fleet-wide series of its Table 1, verified every one of the 78 value cells against the report text extracted with *pdftotext -layout*, and committed them with checksums. Nothing in that dataset is modelled, interpolated or imputed, and a cell absent from the report is absent here.

An earlier version of this subsection claimed that this was the only public series pairing both quantities. That is false, and the correction matters more to the paper than the original claim did. A systematic search found at least six sources pairing cannibalization with supply-driven or total unavailability for the same fleets over the same periods, one of them from the same GAO office in the same year [9]. The right statement is narrower and still worth making: public *per-type* pairing is rare, and it has become rarer, because modern readiness reporting withholds the per-type cut — GAO-21-101SP [10] states that mission capable and availability rates "were deemed to be sensitive" and omits them. The scarcity is a reporting choice of the last two decades, not a fact about the measurement.

The practical consequence is that the estimate below can be replicated out of sample, and Sections 6 and 7 do so twice. The second dataset is Air Combat Command's own standards-and-performance tables as transcribed in [11]: cannibalizations per 100 sorties and mission capable rate, by mission category, for fiscal years 1980 through 1999. The third is [9] itself, the report that falsified the original claim: its enclosure charts the supply-driven rate and the cannibalization rate monthly for three airlift and tanker types, which is 138 paired observations against this panel's 16. Each is weaker than the GAO tables in its own specific ways, and those are stated where each is used.

4. What the data can and cannot identify

It cannot calibrate the model. Cannibalizations per 100 flying hours and a share of aircraft have different denominators, and neither composes into $b(x)$, which is a share of the *disabled fraction*. Recovering β_0 and κ in the model's units would need the fleet counts and flying hours behind the published rates, and the report does not publish them. We therefore offer no calibration, and the companion repository carries a test that fails if anyone adds one.

What the data identifies is the sign, and its significance. That is the quantity the criteria need.

Four properties bound the claim, and all four are stated in the report itself.

SPECIFICATION	SLOPE	STD. ERROR	<i>T</i>	<i>R</i>	<i>N</i>
Within-type, two independent series	+1.089	0.227	+4.79	+0.799	16
Within-type, fleet-wide series added (not independent)	+1.075	0.186	+5.77	+0.791	24
Fleet-wide series alone	+0.179	0.452	+0.40	+0.159	8

Slope is cannibalizations per 100 flying hours per percentage point of supply-driven unavailability. Adding the fleet-wide series shrinks the standard error without moving the estimate, which is the signature of double-counting rather than of more evidence, and is why that row is reported but not used.

Underreporting, up to half. The report cites a 1998 Navy study group finding that as much as 50 per cent of cannibalizations went unrecorded. If reporting discipline falls when a unit is under supply stress — the natural direction — the measured slope is biased *toward zero*, and a positive estimate is conservative. That is the useful direction for a bias to run, and we would be far more cautious about this estimate if it ran the other way.

No direction of causation. Regressing either variable on the other gives a positive slope. Cannibalization plausibly causes unavailability as well, since a donor goes down awaiting its own replacement part, and the model asserts both directions. We claim co-movement, not causation.

Operational tempo moves both. The report attributes the EA-6B's fiscal 1999 rise to its use over Kosovo. Normalizing cannibalizations by flying hours absorbs part of that but not the intensity of use.

Eight years, two independent type series. The sample is small, Navy-only, and fixed by what exists in public rather than chosen. For later years the per-type cut is withheld: GAO's cannibalization reporting largely stops after fiscal 2000, and GAO-21-101SP omits per-type availability as sensitive. Earlier and elsewhere, other pairings do exist, which is what Sections 6 and 7 use — and the second of those is monthly rather than annual, so this bound on the sample is a bound on *this* panel and not on the evidence as a whole.

5. The estimate

The estimator is the within-type slope. Demean both series by aircraft type and pool, so every level difference between types — fleet size, airframe age, mission, reporting culture — drops out, and what remains asks only whether a type cannibalized more in its own worse-supply years. That is precisely the comparative static the model makes.

The F-14 series pairs the all-models cannibalization rate with the unweighted mean of the A, B and D availability rates, because the report publishes no fleet sizes and no weighted mean is available. The fleet-wide series is excluded from the panel because it *contains* the two type series, so including it would double-count them.

That *t* does not survive removing a coincident trend, and an earlier version of this subsection presented it as though it would. The within-type estimator removes type *levels*, not type *trends*. Both types trend over these eight years, in opposite directions, and within each type the two series trend together.

Decomposing each type's identifying covariance into the part generated by its fitted common linear trend and the part from de-

trended deviations:

	S_{xy}	FROM THE COMMON TREND	SHARE OF THE FE WEIGHT
EA-6B	18.12	6.67 (36.8 %)	23.9 %
F-14	63.90	61.66 (96.5 %)	76.1 %
pooled	82.01	68.33 (83.3 %)	—

So 83 per cent of what identifies the slope is coincident trend, and three quarters of the weight sits on the series that is 96.5 per

cent trend. Allow each unit its own linear trend, or first-difference, and the significance goes:

SPECIFICATION	SLOPE	STD. ERROR	t	95 % CI
Within-type, levels (above)	+1.089	0.227	+4.79	[+0.598, +1.580]
Two-way fixed effects, type and year	+1.061	0.268	+3.96	—
Within-type plus unit linear trends	+0.740	0.482	+1.53	[-0.321, +1.802]
First differences, with a constant	+0.550	0.390	+1.41	[-0.299, +1.399]

Detrended, the F-14 contributes nothing (+0.438, $t = +0.37$) and what is left rests on the EA-6B (+0.856, $t = +2.16$ on 5 degrees of freedom) — which does not clear five per cent either, because removing a unit's own trend from an eight-year series leaves five degrees of freedom and the critical value is 2.571. An earlier version of this subsection quoted these as +2.37 and +0.40 against six degrees of freedom and called the EA-6B significant. It is not: **once each type is allowed its own trend, neither type is individually significant**, and the detrended pooled estimate rests on a unit that is itself indistinguishable from zero. A corroborating diagnostic: for the F-14 the correlation with cannibalization is +0.805 contemporaneously, +0.826 at one year's lag and +0.827 at one year's lead. A relation equally strong forwards and backwards in time is the signature of two independently trending series. For the EA-6B the lagged correlations collapse to -0.228 and +0.131 while the contemporaneous one holds at +0.781, which is what a contemporaneous relation looks like.

And at two units, no valid cluster-robust inference exists. Not imprecise: unavailable. The wild cluster bootstrap with Rademacher weights admits four draws and two distinct values of $|t^*|$, so its smallest attainable p-value is 0.500 whatever the data say. Cluster-robust errors fail in the other direction, returning $t = +39.5$ because the two cluster scores nearly cancel and there is no sampling variation left to estimate. What is available is randomization inference, and it has to be read more carefully than an earlier version of this subsection read it. A circular shift preserves a series' marginal distribution and its autocorrelation, but it does **not** preserve its linear trend: averaged over the full set of rotations the fitted trend slope is exactly zero, for every unit of every panel in this section. The levels reference set is therefore trendless *by construction*, so ranking a trending observed

slope against it asks whether the two series trend together — which is the thing 83.3 per cent of this panel's identifying covariance already is, not an independent check on it. **We withdraw the levels p-value** ($p = 0.016$, rank 1 of 64); for the F-14, whose series is 96.5 per cent trend, the observed trend slope is itself the most extreme value in its own reference set, which is what that degeneracy looks like from the inside. Detrended, the observed statistic and the reference set have had the same unit trends removed, the comparison is like for like, and **$p = 0.047$ (rank 3 of 64) stands**. A naive within-type year permutation gives $p = 0.0005$ and we do not quote it, because destroying the serial dependence makes it anti-conservative.

The sign is positive in every specification we ran. Its significance holds only in levels. Point estimates range from +0.44 to +1.19 across more than twenty variants, both types are positive separately, and no single observation or year drives it: leave-one-out gives +0.971 to +1.186 with t never below +3.44. Two attacks we expected to work did not: residual autocorrelation is *negative* on this panel (pooled -0.303 in levels and -0.233 detrended, Durbin-Watson 2.568), and the two types give nearly identical separate slopes in levels (+1.008 at $t = +3.06$ and +1.115 at $t = +3.32$, six degrees of freedom each).

What must not be done with that diagnostic is generalize it, and an earlier version of this subsection did. It reported that every robust standard error is smaller than the one reported and concluded that the reported error is therefore conservative. Both statements were measured on this panel and hold on this panel only. Eight observations per unit is in any case too short a series for a heteroskedasticity-and-autocorrelation-consistent estimator to say anything: a Bartlett bandwidth of four spans half the series and one of twelve exceeds it outright, and the Newey-West sta-

tistics this panel returns under those bandwidths — $+14.70$ at lag 4 and $+24.46$ at lag 12, against the reported $+4.79$ — are the signature of a bandwidth collapsing on a series too short to carry it, not evidence for anything. They are quoted here to record that they must not be quoted as support. Section 7's panel is the only one of the three long enough to carry the question honestly, and there the answer reverses. What survives without qualification is narrower: on this panel the reported standard error is not flattered by serial correlation, and what it is not is a defence against the trend — and the trend is what carries the t .

The qualitative assumption of *A Curvature Criterion for Cannibalization Feedback* — the one that makes the feedback positive and a fold possible at all — therefore holds in sign on this data, and the strength of that conclusion depends on a replication rather than on this panel. Sections 6 and 7 supply two.

Our simulator's negative sign is therefore a property of its donor policies, not of cannibalization. The simulation study of *Readiness at the Critical Point* found the propensity falling in the disabled fraction, because a pull costs a flyer only when no already-down airframe holds the needed part, and that pool grows as the fleet degrades. That mechanism is real and the measurement of it stands. What an earlier version of that paper drew from it — that cannibalization is self-limiting, full stop — generalized from one simulator's policy set to the phenomenon. Real fleets, under whatever policies they actually run, move the other way. The two findings are compatible and the disagreement is informative: it localizes to the policy, which is a modelling choice an organization can change, and not to the mechanism, which it cannot.

6. Replication on independent Air Force data, which survives the same attack

If the Navy panel's significance is a trend artefact, the way to find out is to run the identical attack on a different service, a different decade and a different denominator. Air Combat Command's own standards-and-performance tables [11] report cannibalizations per 100 sorties and mission capable rate by mission category — Fighters for fiscal years 1980 through 1999, Bombers and Special Mission for 1990 through 1999. Dropping the two category-years with no cannibalization value leaves 38 paired observations, more than twice the Navy panel.

It is weaker evidence than the GAO tables in four specific ways, and all four belong here rather than in a footnote. The primary document is not online; the values come from a transcription by a research institute [11] and we have not checked them against the report itself, where every GAO cell was checked against the report's own text. The transcribed series runs past its source's stated range of FY89 to FY98, which cannot be right, and we have no explanation for it. The availability measure is mission capable rate, so our regressor is $100 - MC$, which is *total* unavailability and includes maintenance causes the model does not attribute to supply — this is a replication on a **proxy**, not a strict one. And the denominator is sorties rather than flying hours, so the two panels must never be pooled; the repository carries a test that fails if anyone writes a function combining them.

With those stated, the result:

SPECIFICATION	SLOPE	STD. ERROR	<i>T</i>	<i>DF</i>
Within-category, levels	+0.757	0.174	+4.36	34
Within-category plus category linear trends	+0.813	0.181	+4.49	31
First differences, with a constant	+1.093	0.282	+3.88	33

The significance survives detrending here, and strengthens. Side by side with the Navy panel under the identical

specifications:

	GAO NAVY, <i>N</i> = 16	ACC AIR FORCE, <i>N</i> = 38
Within-unit, levels	+1.089 (<i>t</i> +4.79)	+0.757 (<i>t</i> +4.36)
Plus unit linear trends	+0.740 (<i>t</i> 1.53)	+0.813 (<i>t</i> +4.49)
First differences	+0.550 (<i>t</i> 1.41)	+1.093 (<i>t</i> +3.88)
Circular-shift <i>p</i> , levels	<i>withdrawn</i>	<i>withdrawn</i>
Circular-shift <i>p</i> , detrended	0.047	0.0006

The reason the two panels behave differently is measurable and is the point. Of the Navy panel's identifying covariance, 83.3 per cent is coincident linear trend. Of the Air Force panel's, **11.2 per cent** is. The Air Force estimate is identified by within-category

variation around each category's own trend, which is the comparative static the model makes, and that is why removing the trend does not remove the result.

Serial correlation does cut this estimate, and it survives. Unlike the Navy panel, the residuals here are positively autocorrelated — pooled $+0.477$ in levels and $+0.359$ detrended — so a heteroskedasticity-and-autocorrelation-consistent error is *larger* than the one reported, in the direction that costs significance. The detrended t falls from $+4.49$ to $+3.21$ at a Bartlett bandwidth of four. That is the correct direction for this diagnostic to run and the result clears five per cent either way. The bandwidth is only defensible for the twenty-year Fighters series, since the

SPECIFICATION	SLOPE	STD. ERROR	<i>T</i>	<i>DF</i>
Within-category, levels	+1.293	0.358	+3.61	22
Within-category plus category linear trends	+2.482	0.354	+7.02	19
First differences, with a constant	+2.017	0.402	+5.02	21

Two of the three strengthen and the levels fit weakens, which is the only reading these numbers support: the levels t falls from $+4.36$ to $+3.61$ because the sample is two thirds smaller, while the two specifications that answer the trend objection roughly double — $+4.49$ to $+7.02$ detrended and $+3.88$ to $+5.02$ differenced. The identifying covariance that is coincident trend falls from 11.2 to 6.3 per cent, the detrended circular-shift p-value is 0.0031 (rank 2 of 640), and the detrended Newey-West t at a bandwidth of four is $+5.07$. So the restriction buys identification rather than precision, and the specifications it buys it in are the ones this section exists to defend. A first draft of this paragraph said every specification strengthens; the test that pins these numbers asserted that and failed on the levels row, which is the correction. We report the full panel as primary because it is the larger sample and because dropping observations that disagree with a source's stated range is a choice a referee should see made rather than assumed, but the objection does not cut the way it appears to: **the unexplained years are not carrying this result**. What the restriction cannot do is explain where they came from, and we still have no explanation.

Two things we are not claiming. This is not independent confirmation in the strict sense, because $100 - MC$ is not the model's π ; a mechanism that moved maintenance-driven unavailability and cannibalization together would produce the same estimate. And one of three categories has the wrong sign in levels: Special Mission gives -0.122 ($t = -0.30$), flipping to $+1.334$ ($t = +3.27$) once its trend is removed. We report that rather than dropping the category, and note that its identifying covariance is near zero ($S_{xy} = -11.0$ against the Fighters' $+600$), so the sign in levels is not well determined either way. Fighters give $+0.484$ ($t = +10.5$, detrended $+0.487$) and Bombers $+3.030$ ($t = +6.4$, detrended $+3.158$).

Across two services, two decades and two denominators, then, cannibalization intensity rises with unavailability, and the sign is robust to every specification either panel admits. The *strength* of that conclusion comes from the Air Force panel, because the

other two categories have nine observations each, which is why no longer lag is quoted for this panel.

The out-of-range years can simply be dropped, and the result improves. One of the four weaknesses above is that the transcription runs past its source's stated range of FY1989 to FY1998. Restricting the panel to exactly that range leaves 26 paired observations — Fighters 10, Bombers 8, Special Mission 8:

Navy panel's strength was trend; and the Air Force panel's weakness is its provenance, because we have not seen its source document. That trade — strong where the data is weak, weak where the data is strong — is what Section 7 was built to break, with a third panel that is eight times the first panel's size and carries the model's own regressor.

7. A third panel, large enough to answer the sample-size objection

The most dangerous objection to Section 5 is not the trend. It is $n = 16$. Eight years, two series, and a result whose significance rests on one of them after detrending is a thin basis for a structural claim, and the Air Force replication of Section 6 answers it only partly: 38 observations is better, but its regressor is a proxy and its primary document is one we have not read.

This subsection adds a panel of **138 paired observations**, monthly, with the supply-driven regressor the model actually means, from a GAO report in the same series as the first. The trade is a different one: for the first time, the numbers were not read off a printed table.

GAO-01-495R [9] — the report whose existence falsified an earlier version's claim that GAO-01-771 was the only public pairing — charts, in its enclosure on pages 14 through 19, both the not-mission-capable-supply rate and cannibalizations per 100 sorties for the C-5, the KC-135 and the KC-10, monthly from November 1996 to September 2000. Six charts, each sourced in the report to "GAO Analysis of U.S. Air Force, Air Mobility Command data", and each plotting the fleet's rate against the Air Mobility Command standard for it. After dropping two months that one of the two charts does not reach, that is 138 paired observations: C-5 46, KC-135 45, KC-10 47.

The charts print no data labels. There is no cell anywhere in that document to check a value against, and that is the central limitation of this panel. What we did instead is read the PDF's vector geometry: the series are real polylines, so the values come from coordinates rather than from sampling an image. The ex-

tractor calibrates each axis by least squares on the positions of the tick labels and refuses to proceed unless five conditions hold,

all five reported on every run:

CHECK	WHAT IT RULES OUT	WORST CASE OVER THE SIX CHARTS
value-axis tick labels collinear	a misread or non-linear axis	0.7 % to 1.4 % of one tick interval
time-axis labels collinear	a misread time axis	0.000 to 0.021 months
every year tick anchored to <i>Oct</i>	a calendar shifted by whole months	24 of 24 ticks
polyline vertices on a monthly grid	the wrong polyline, or non-monthly data	0.032 to 0.079 months
the two series of one aircraft overlap	two charts read on different grids	45 to 47 of 47 months shared

The grid check is the discriminating one. Each chart contains a frame of 114 to 122 vertices and a copy of the series closed down to the axis baseline, as well as the series itself at 45 to 47 vertices; the frames sit 0.19 to 0.25 months off the monthly grid and are rejected by it, and a series that was not monthly would be off by about 0.5. So a month is identified to within about two and a half days and a value to within about one per cent of a tick. **That bounds the digitisation error. It does not establish that the chart is faithful to Air Mobility Command's records** — no procedure applied to *this* document could, and Section 8 reaches outside it for one that can.

Four further properties of the extraction are worth stating because each of them was a defect first. Colour does not discriminate the data line — only two of the six charts draw it in grey, the others use the frame's black — so the series is selected structurally, as the shortest polyline landing on a monthly grid, which also excludes the Air Mobility Command standard line by its being far shorter than any monthly series. Tick clustering has to require even spacing, because on one chart the page number 19 sits in the same horizontal band as the ticks and threw that chart's fitted scale by a factor of five. The months with no vertex are carried as gaps rather than interpolated, which is why the first-difference specification below has $n = 135$. And the two charts for one aircraft do *not* plot identical month grids — they cover 45 to 47 months of the same 47-month axis — so the rows are an inner join on the month index, and an unmatched month is dropped rather than paired against the wrong one.

Two weaknesses in that extractor are narrower than the checks above make them sound, and an audit of the code found both. The structural selection works, but not by a wide margin: each chart also contains a copy of the series closed down

to the axis baseline, at 50 or 51 vertices against the data line's 45 to 47, and that copy **passes every one of the five checks above**. It is excluded only by being four or five vertices longer. Its off-grid error is sometimes smaller than the chosen line's, so none of the reported diagnostics discriminates between them; what does is a vertex count, and on four of the six charts there are two tied candidates. The discriminating property that is *not* checked is the obvious one — the baseline copy's minimum value is about 0.02 where the data lines' minima run 1.90 to 32.30 — and the extractor now asserts it. Separately, the tick-clustering step keeps the largest evenly spaced cluster and silently discards the others, which on the C-5 chart throws away two genuine axis ticks (0 and 5) that bucket with the page number; the reported residual is computed over the survivors, so a discard was invisible. Refitting that chart with all seven printed ticks moves its values by at most 0.02 percentage points, so the committed table is unaffected, but the check was weaker than its description.

The calendar is fixed, and an earlier draft of this subsection said it was not. The time axis labels every second month, *Oct-96*, *Dec*, *Feb*, and so on, not only the fiscal years; the year numbers are the second line of an *Oct-<yy>* label on all twenty-four ticks across the six charts, so month index m is October 1996 plus m months. The parity is pinned by the same labels: they sit at even indices while the series occupy 1 through 47, so a reading that put the first point at October 1996 would put the point at index 2 in November, where the printed label says December. Nothing in the estimate depends on this, since every specification is within-unit and a common shift cancels, but a dataset that can be dated should be.

One independent check on the value calibration exists, and it passes:

	MONTHS	MEAN NMCS, PER CENT	MEAN CANNIBALIZATIONS PER 100 SORTIES
C-5	1996-11 to 2000-08	18.07	52.59
KC-135	1996-11 to 2000-08	13.11	8.68
KC-10	1996-11 to 2000-09	4.25	3.61

The C-5 was this period's notorious airlift maintenance problem and the report is largely a complaint about it; the KC-10 was the newest and most reliable of the three. An axis error of the kind

the page number caused would not preserve that ordering or those magnitudes.

With all of that stated, the estimate:

SPECIFICATION	SLOPE	STD. ERROR	<i>T</i>	<i>DF</i>	95 % CI
Within-aircraft, levels	+0.575	0.129	+4.47	134	[+0.320, +0.829]
Within-aircraft plus aircraft linear trends	+0.500	0.139	+3.59	131	[+0.225, +0.776]
First differences, with a constant	+0.252	0.152	+1.65	133	[-0.049, +0.553]

The sign survives the attack that killed the Navy estimate. Only 25.3 per cent of the identifying covariance here is coincident linear trend, against 83.3 per cent on the Navy panel, and circular-shift randomization over 97 290 shifts ranks the observed slope 241st detrended ($p = 0.0025$). The levels ranking is withdrawn for the reason Section 5 gives — the rotation set is trendless by construction, so in levels it tests trend coincidence rather than the slope — and on this panel the detrended p-value was the one worth having anyway. All three aircraft are positive separately in levels — C-5 +1.531 ($t = +3.96$), KC-135

+0.262 ($t = +2.83$), KC-10 +0.266 ($t = +1.90$) — and all three remain positive detrended.

This is the panel where serial correlation can actually be measured, and it costs the detrended fit most of its margin. Forty-five to forty-seven monthly observations per unit is the only series length in this paper that a heteroskedasticity-and-autocorrelation-consistent estimator can be run on without the bandwidth swallowing the sample. Run on it, the residuals are positively autocorrelated — pooled +0.305 in levels and +0.301 detrended — and the correction runs the opposite way to the Navy panel's:

SPECIFICATION	REPORTED <i>T</i>	NEWKEY-WEST, LAG 4	NEWKEY-WEST, LAG 12
Within-aircraft, levels	+4.47	+2.44	+2.17
Within-aircraft plus aircraft linear trends	+3.59	+2.18	+1.90

The levels fit clears five per cent under either bandwidth. The detrended fit clears it at a bandwidth of four and **does not** clear it at twelve, where +1.90 sits just under the 1.978 critical value at 131 degrees of freedom. A year of lags on monthly data is the more defensible choice of the two. So the honest reading of this panel's detrended specification is that it is **marginal once monthly persistence is accounted for**, not comfortable, and the three-panel summary below is stated under independent errors because that is the only basis on which all three are comparable.

Nor is cluster-robust inference available here. Section 5 argues that two clusters admit none; three is not meaningfully better — the wild cluster bootstrap with Rademacher weights admits eight draws, and asymptotics in the number of clusters have nothing to work with at $G = 3$. **No panel in this section supports valid inference under arbitrary within-unit dependence.** What the three panels jointly support is a sign, repeated across indepen-

dent data, which is a weaker and more defensible claim than any one t statistic in them.

It does not survive first-differencing, and that belongs in the main table rather than a footnote. Two things about that. A month-on-month change in either series is mostly noise, so first-differencing at monthly frequency is a more severe test than the same specification on an annual panel, and failing it is weaker evidence against than failing the detrended fit. But the KC-135 carries 71.7 per cent of the fixed-effects weight and 55.9 per cent of its own identifying covariance is trend — it is, within this panel, the unit that resembles the Navy panel's problem — and detrended its slope falls to +0.137 at $t = +1.53$. The panel's strength is not uniform across its units.

What the three panels do and do not jointly establish. Each is weak in a different place, and the places do not overlap:

	GAO NAVY, <i>N</i> = 16	ACC AIR FORCE, <i>N</i> = 38	GAO AIRLIFT, <i>N</i> = 138
unit	aircraft type	mission category	aircraft type
regressor	NMCS rate	100 – MC, a proxy	NMCS rate
outcome denominator	per 100 flying hours	per 100 sorties	per 100 sorties
provenance	table checked against the report	transcription, source unread	charts, digitised

	GAO NAVY, <i>N</i> = 16	ACC AIR FORCE, <i>N</i> = 38	GAO AIRLIFT, <i>N</i> = 138
frequency	annual	annual	monthly
levels	+1.089 (<i>t</i> +4.79)	+0.757 (<i>t</i> +4.36)	+0.575 (<i>t</i> +4.47)
plus unit trends	+0.740 (<i>t</i> 1.53)	+0.813 (<i>t</i> +4.49)	+0.500 (<i>t</i> +3.59)
first differences	+0.550 (<i>t</i> 1.41)	+1.093 (<i>t</i> +3.88)	+0.252 (<i>t</i> 1.65)
identifying covariance that is trend	83.3 %	11.2 %	25.3 %

No two of them measure the same pair of quantities, so none of the slopes may be compared for magnitude and no two may be pooled. The airlift panel is the one that looks poolable and is not: it shares its regressor with the Navy panel and its denominator with the Air Force panel. The repository carries a test that fails if any two are combined, and the three cannibalization columns are named differently on purpose so that a natural join cannot line up what must not be joined.

What they share is a sign, positive in all nine specifications across the three, and two of the three survive removing unit-specific linear trends. The most compact honest statement of what this paper establishes is those nine slopes side by side: nine positive slopes, **six** of them clearing five per cent under independent errors, and the three that do not named rather than omitted. An earlier version of this sentence said seven cleared and three did not, which is ten specifications and wrong twice over; the count is six and three. The honest summary of the three together is that **the sign is as well established as public data allows, and no magnitude is established at all**: the Navy panel's magnitude is not identified, the Air Force panel's regressor is not the model's, and the airlift panel's values were read off chart geometry rather than a printed table. A referee who distrusts any one of the three is left with two that fail somewhere else.

8. A fourth report, which checks the other three and closes one confound

Everything above rests on three tables that cannot check each other: they share no variable pair, so no two of them can even be compared for magnitude. This subsection reaches for a fourth source that is not a panel and cannot be estimated on, and uses it for the three things the panels cannot do for themselves.

	DIGITISED FY2000 MEAN	PRINTED, INVENTORY-WEIGHTED	GAP	PRINTED SUB-MODEL RANGE
C-5	51.37	48.98	+4.9 %	10.5 to 54.6
KC-135	10.65	9.59	+11.0 %	7.8 to 10.3

These are not the same quantity and the comparison should not be read as though they were. One is a mean over the eleven fiscal-2000 months the panel reaches, taken off a chart; the other is an inventory-weighted combination of sub-model rates using a

GAO-02-86 [1] is the GAO's dedicated report on cannibalization — the same report Section 1 cites for the practice's prevalence, from the same office nine months after GAO-01-771. Its appendices tabulate the cannibalization rate, active inventory and average age of 91 aircraft type/model/series across four services. **The three columns are not dated alike, and the two appendices do not date them alike either**, which matters for how any of them may be used: the rate is fiscal year 2000 in both tables, average age is March 2001 in both, and active inventory is March 2000 in the Navy and Marine Corps table but March 2001 in the Air Force one. Those are the printed column headings, and the dataset carries them. **It carries no unavailability measure at all**, so it reports one side of the law and no slope in this paper is estimated on it; the repository's loader has no entry point that accepts it as a panel. Unlike the digitised charts it is printed text, so every value is checkable, and all 259 value cells were — 231 by cross-checking two independent PDF extractors against each other and the remaining 28 by hand, because the 14 rows whose rate the source gives as a footnote marker rather than a number do not render in the layout view. Zero mismatches.

First: the one cell two GAO reports both print, agrees exactly. GAO-01-771's Table 7 and GAO-02-86's Table 1 each give the EA-6B's fiscal 2000 cannibalization rate as 16.7 per 100 flying hours. One cell is not a validation of either table, but it is the only direct cross-report comparison available, and a transcription error in either file would have to be shared to survive it.

Second, and the reason this subsection exists: the digitised panel has an external check after all. Section 7 established that nothing *inside* GAO-01-495R can check a value, and an earlier draft of it overstated that into a claim that no procedure could. GAO-02-86 prints the fiscal 2000 cannibalization rate per 100 sorties for the C-5 and the KC-135 — the same denominator as the airlift panel:

March 2001 snapshot, and two reports may compute a rate differently. **So this bounds the scale and rules out a calibration error; it validates no individual month.** Ruling out a calibration error is worth having: the failure that nearly shipped in this

extraction was a page number clustered in with the axis ticks, which threw one chart's fitted scale by a factor of five, and nothing internal to the document would have caught it. The C-5 agreement is good; the KC-135's is adequate and sits just above the top of its own sub-model range, so if either series runs slightly high it is that one.

SPECIFICATION	UNWEIGHTED (REPORTED)	INVENTORY-WEIGHTED
Within-type, levels	+1.089 (<i>t</i> +4.79)	+1.064 (<i>t</i> +4.74)
Plus unit linear trends	+0.740 (<i>t</i> 1.53)	+0.714 (<i>t</i> 1.47)
First differences	+0.550 (<i>t</i> 1.41)	+0.552 (<i>t</i> 1.41)
identifying covariance that is trend	83.3 %	84.1 %

Every verdict holds and the slope moves by about two per cent. We keep the unweighted mean, because one March 2000 snapshot applied to a panel starting in 1993 — across which the F-14A was being drawn down — trades one known error for another; what has changed is that this is a measured choice rather than an assumption. One incidental sign that GAO itself aggregated by something close to fleet size: its own all-models F-14

Third: an assumption in Section 5 can now be tested, and it does not carry the result. That panel combines the F-14A, F-14B and F-14D availability rates unweighted, and the reason recorded in both the paper and the code was that the report publishes no fleet sizes. GAO-02-86 publishes them — 87, 71 and 46 as of March 2000. Applying them:

figure for fiscal 2000 is **16.2**, against **17.25** for the inventory-weighted combination of its per-model rates and **19.30** for the unweighted one.

Fourth, and the one new fact rather than a check: average fleet age does not predict cannibalization. Regressed within service, because the two services' rates have different denominators:

SERVICE	<i>N</i>	SLOPE PER YEAR OF AGE	SE	<i>T</i>	<i>R</i>
Navy and Marine Corps	49	−0.070	0.083	−0.84	−0.122
Air Force	28	−0.067	0.310	−0.22	−0.042

Nothing in either service, and faintly negative in both. The thirty-year-old C-5A cannibalizes at 54.6 per 100 sorties while the forty-two-year-old KC-135E does so at 7.8.

What that does and does not bound, stated as an interval rather than as an absence. These samples are small and the rates are dispersed, so failing to find an effect is not the same as bounding one near zero. Across the age range each service's table actually spans, the 95 per cent interval on the age coefficient admits a movement of **−1.41** to **+0.58** standard deviations of the cannibalization rate in the Navy and Marine Corps table, and **−1.58** to **+1.28** in the Air Force one. An age effect large enough to matter is therefore **not excluded**; what is excluded is the specific shape the confound needs, a *positive* association strong enough to generate the panels' sign, for which the point estimates run the wrong way in both services and the upper limits are reached only at the far end of a forty-year span.

That closes a confound the paper would otherwise owe an answer on. A reader can object that the positive sign of Sections 5 through 7 is an artefact of airframe age — that old aircraft are both more often down for parts and more often robbed, with no causal link between the two. If age drove cannibalization, it should appear in a 91-type cross-section, and we cannot detect it — which, as the interval above says, is weaker than saying it is not there. An earlier version of this subsection said it "does not

appear at all", which is exactly the inference Section 9 warns against in the aggregation case: an estimate with wide intervals reporting nothing is a statement about power. This is also a cross-sectional association with no unavailability control, so it is not evidence *about* the model's law; it is evidence against one specific alternative explanation of it, which is a narrower and more useful thing.

9. What aggregation destroys, which is power and not the slope

The fleet-wide Navy series finds nothing: slope **+0.179**, *t* = **0.40**. An earlier version read that as the estimate changing under aggregation and said an analyst holding only the aggregate would be "wrong about the slope by a factor of six". That is not supportable and a referee would stop there.

The two estimates are not significantly different. The difference is **+0.910** with a standard error of **0.506** under independence, so *t* = **+1.80**, two-sided *p* ≈ **0.088** — and independence overstates the separation, since the aggregate *contains* the two types. More decisively, the aggregate's own 95 per cent interval is **[−0.927, +1.285]**, which **contains the within-type point estimate of +1.089**. The aggregate series is not evidence of a flat slope. It is evidence of nothing in either direction, and its half-width exceeds the entire claimed effect.

What aggregation destroys is the identifying variance, and that part is real, mechanical, and worth the subsection:

SERIES	S_{xx}	$SD(X)$	RANGE OF X
EA-6B	17.98	1.60 pp	5.1 pp
F-14	57.32	2.86 pp	8.9 pp
Navy aggregate	1.19	0.41 pp	1.0 pp

Sixty-three times less. The aggregate measure of supply-driven unavailability moves one percentage point across eight years, because the EA-6B rises while the F-14 falls — their correlation is -0.519 — so averaging cancels the variation that identifies the law.

The operational consequence survives in this weaker and more precise form. The quantity the criteria of *A Curvature Criterion for Cannibalization Feedback* need is $2 \beta_0 \kappa$, a within-unit slope. Readiness is published and argued about at the service level, and at that level the estimate has a confidence interval spanning the entire plausible effect range, so it cannot test the law at all. That is a statement about statistical power and about where readiness statistics are published. It is not a claim that the slope changes under aggregation, and the data cannot support the stronger version.

10. Limitations

The limitations of each panel are stated where that panel is used rather than collected here, and the four that bound the estimate most are worth repeating. No panel has enough clusters for cluster-robust inference: two on the Navy panel, where it does not exist rather than being imprecise, and three on each of the others, which is not meaningfully better. Nor does any panel but the airlift one have series long enough to support a heteroskedasticity-and-autocorrelation-consistent error, and on that one the correction leaves the detrended fit marginal. **Every t statistic above assumes independent errors**, which is the only assumption under which all three panels are comparable and is not defensible on all three; what this paper establishes is a sign repeated across independent data rather than any one of those statistics. The Air Force panel rests on a transcription of a document we have not read, 12 of its 38 observations fall outside the range its cited source can cover — concentrated in the category that supplies most of the identifying variance — and its regressor is total unavailability rather than the supply-driven rate the model means, so it is a replication on a proxy. Of those three, only the provenance is unanswerable: Section 6 reports the panel restricted to exactly the range its source states, on which every specification strengthens, so the unexplained years are not what carries the estimate. The airlift panel of Section 7 answers the first panel's sample size and the second panel's proxy, and pays for it in provenance: its 138 values were digitised from the vector geometry of charts that print no numbers. Section 8 reaches

outside that document for the only external check available and finds the annual means within five and eleven per cent of an independently printed table, which bounds the scale and rules out a calibration error — and validates no individual month. **No panel here is strong in every way, and the argument rests on their weaknesses not coinciding rather than on any one of them being sound.**

No fleet experiment of our own. Every operational number here comes from statistics someone else published; we have run nothing on an actual fleet.

11. Conclusion

Three panels of published fleet statistics, measuring different pairs of quantities on different services in different decades, agree on one thing: within a unit, cannibalization intensity rises with unavailability. The sign is positive in all nine specifications the three panels admit, six of those clear five per cent under independent errors, and two of the three panels keep their significance once each unit is allowed its own linear trend. That is the whole of the claim. No magnitude is established at all — the Navy panel's magnitude is not identified, the Air Force panel's regressor is total rather than supply-driven unavailability, and the airlift panel's values were read off chart geometry rather than a printed table — and no panel identifies a direction of causation, because cannibalization plausibly causes unavailability as well.

What the three panels buy is not precision but non-coinciding weakness. Each fails somewhere the others do not, and the argument rests on that rather than on any one t statistic in it; a reader who distrusts any one of the three is left with two that fail somewhere else. The one bias here that is documented rather than assumed runs in the conservative direction: the Navy report cites a 1998 study group finding that as much as 50 per cent of cannibalizations went unrecorded, and if reporting discipline falls when a unit is under supply stress, the measured slope is biased toward zero.

Two things would settle what this paper cannot. Records that attribute each grounded airframe either to its own failure or to someone else's part would identify the propensity directly, instead of through a published rate whose denominator is flying hours or sorties; that is a reporting discipline rather than a modelling advance, and it is the thing the GAO found half-missing in 1998. And the evidence available in public is getting thinner rather than thicker, because modern readiness reporting with-

holds the per-type cut [10]. The scarcity of pairings like these three is a reporting choice of the last two decades, not a fact

about the measurement.

References

- [1] U.S. General Accounting Office (now the Government Accountability Office). *Military aircraft: services need strategies to reduce cannibalizations*. GAO-02-86, 2001. www.gao.gov/products/gao-02-86
- [2] C. C. Sherbrooke. METRIC: a multi-echelon technique for recoverable item control. *Operations Research*, 16(1):122–141, 1968. doi.org/10.1287/opre.16.1.122
- [3] W. W. Fisher, J. J. Brennan. The performance of cannibalization policies in a maintenance system with spares, repair, and resource constraints. *Naval Research Logistics Quarterly*, 33(1):1–15, 1986. doi.org/10.1002/nav.3800330102
- [4] D. P. Gaver, K. E. Isaacson, J. B. Abell. *Estimating aircraft recoverable spares requirements with cannibalization of designated items*. RAND Corporation, R-4213-AF, 1993. www.rand.org/pubs/reports/R4213.html
- [5] J. Sheng, D. Prescott. A hierarchical coloured Petri net model of fleet maintenance with cannibalisation. *Reliability Engineering and System Safety*, 168:290–305, 2017. doi.org/10.1016/j.ress.2017.05.043
- [6] J. Sheng, D. Prescott. Using a novel hierarchical coloured Petri net to model and optimise fleet spare inventory, cannibalisation and preventive maintenance. *Reliability Engineering and System Safety*, 191:106579, 2019. doi.org/10.1016/j.ress.2019.106579
- [7] I. Dobson, B. A. Carreras, V. E. Lynch, D. E. Newman. Complex systems analysis of series of blackouts: cascading failure, critical points, and self-organization. *Chaos*, 17(2):026103, 2007. doi.org/10.1063/1.2737822
- [8] U.S. General Accounting Office (now the Government Accountability Office). *Navy inventory: parts shortages are impacting operations and maintenance effectiveness*. GAO-01-771, 2001. www.gao.gov/products/gao-01-771
- [9] U.S. General Accounting Office (now the Government Accountability Office). *Military readiness: updated readiness status of U.S. air transport capability*. GAO-01-495R, 2001. www.gao.gov/products/gao-01-495r
- [10] U.S. Government Accountability Office. *Weapon system sustainment: aircraft mission capable rates generally did not meet goals*. GAO-21-101SP, 2020. www.gao.gov/products/gao-21-101sp
- [11] Air Combat Command, Director of Logistics. *Ten years look back: standards and performance FY89-FY98*. Langley AFB, November 1998. Tables 4 and 5, as transcribed in Project on Defense Alternatives, *The readiness crisis of the U.S. Air Force*. www.comw.org/pda/afread02.html
- [12] R. P. Trost, M. Ye, J. Hoover, S. C. Goodwyn, J. M. Jondrow. The relationship of weapon cannibalization to weapon system mission capability. *Military Operations Research*, 9(2):31–42, 2004. doi.org/10.5711/morj.9.2.31