Metrics of Research Impact in Astronomy

John Körmendy

Department of Astronomy, University of Texas at Austin, USA

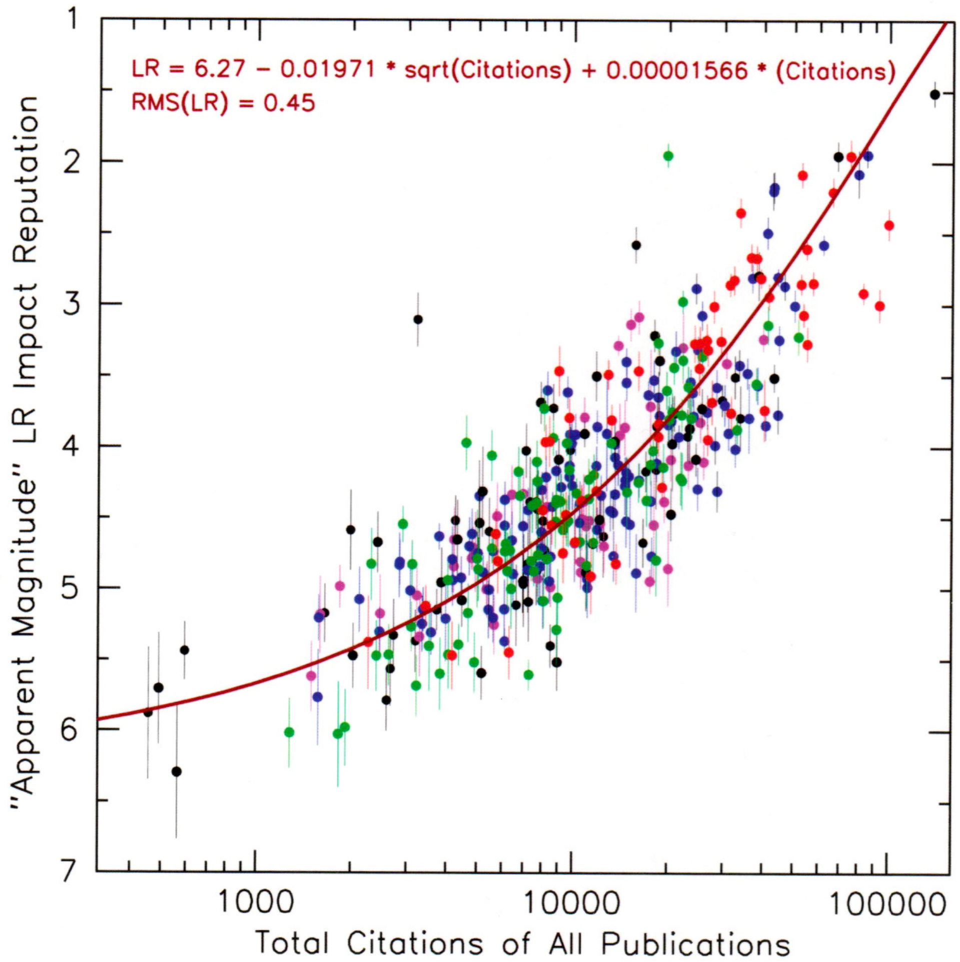
Max-Planck-Institute for Extraterrestrial Physics,
Garching-by-Munich, Germany

Aim:
<u>to measure career research impact using 10 metrics</u>
that are easy to get via SAO/NASA ADS and
<u>whose interpretation is calibrated using</u>
"LR" = the mean of "votes" from 22 experienced astronomers
on the research impact of 510 faculty members
at 17 highly-ranked universities worldwide.

Astronomical Society Of The Pacific
CONFERENCE SERIES

https://myasp.astrosociety.org/product/CS530/vol-530-monograph-8-research-of-metrics-impact-in-astronomy

METRICS OF RESEARCH IMPACT IN ASTRONOMY



John Kormendy

Why do we care?

We benefit from quantitative tools to measure success:

Students & postdocs benefit from calibration of standards needed to start careers.

Resource committees (hiring, tenure, money, prizes, …) benefit from reliable tools to measure impact.

At all ages, the results help to guide decisions on <u>what questions to work on</u> and tell us <u>how results are received</u>.

I hope that the use of metrics will make us more fair in our attribution of credit for discoveries.

How does my work relate to Hungarian science's use of mtmt.hu?

Metrics used here are from the NASA/SAO Astrophysics Data System "ADS" – https://ui.adsabs.harvard.edu : very complete for astronomy and related fields.

All calibration here is specific to astronomy.

Calibration will be different in fields that are far from astronomy.

The best metrics may be different in different fields.

Citation behavior (e. g. who is first author?) may be different.

ADS is carefully curated and reliable.

Web of Science is similarly reliable.

Google scholar is NOT carefully curated and contains many mistakes:
It overcounts citations wrt ADS by a mean of 26 % (range = 5 % – 40 %)

I am not familiar with other sources of metric data.

I hope that my book shows that metrics provide reliable information.

My most important result may be to demonstrate how <u>normalized citations</u> allow reliable comparison of big-team and non-big-team people.

Problem

Many decisions (job hires, tenure, …),
are uncomforably based on qualitative opinions.

We never do research
that is so strongly based on personal opinions.

My book tries to lend to such judgment processes
some of the quantitative rigor that we use
when we do research.

“Wholistic” decisions are based on many judgments,
not just research. I focus only on research.

Problem 2: Metrics are often quoted “in a vacuum” without
a comparison sample. I provide a robust comparison sample.

Study sample @ epoch 2017.0 =
510 faculty members at 17 institutions worldwide.

1 – Caltech
2 – Harvard University
3 – University of California at Berkeley
4 – University of California at Santa Cruz
5 – Princeton University
6 – University of Arizona

Emphasize high ranking,
here US News and World Report 2016 ranking for space sciences

8 – University of Chicago
9 – University of Cambridge
10 – University of Texas at Austin
10 – University of Toronto

28 – University of Michigan, Ann Arbor

12 – University of Oxford

55 – Australian National University

15 – Leiden University

18 – Pennsylvania State University, University Park

? – University of Hawaii
? – University of Groningen

Emphasize
job ladders similar to those in the USA
(hence, e. g., not Germany)

ADS metrics are easy to get.
The challenge is interpretation.
To calibrate interpretation, develop metric "LR":

LR estimates how much impact a person's research
has had on clientele communities
<u>as perceived by those clientele communities</u>.

How much "mental resolution" do we need? Suggest:

We need 2 – 3 steps above and 2 – 3 steps below the mode
plus a tail at highest impact.

$\Rightarrow$ LR is similar to "apparent magnitudes" of stars.

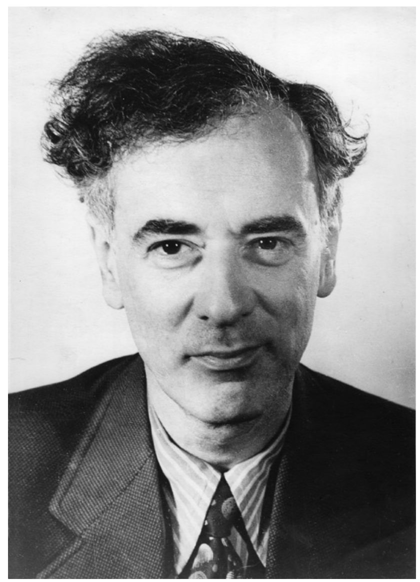
The Landau Scale

0.0: Newton

0.5: Einstein

1 : Top Nobel Prize winners: Bohr, Dirac, Fermi, Heisenberg, Schrödinger, …

2 : Typical Nobel Prize winners or equivalent



Adapted from an
idea by Lev Landau

The Landau-Richter Scale

For us, this is the "tail at highest impact".

This is an impact metric.

LR = 0.0: Newton

LR = 0.5: Einstein

LR = 1 : Top Nobel Prize winners: Bohr, Dirac, Fermi, Heisenberg, Schrödinger, …

LR = 2 : Typical Nobel Prize winners or equivalent

LR = 3 : Top owners of the state of the art in their field (e. g., National prize winners)

LR = 4 : Intermediate impact

LR = 5 : Normal successful career ≈ mode of distribution

LR = 6 : Intermediate impact

LR = 7 : Low research impact (often because impact is in other areas, e. g., teaching)

LR = 8 : No research

LR measures people's research impact, not likeability
or teaching ability
or non-research service
or even intelligence ….

hence "LR".

I invited 42 men and 12 women →

22 LR Evaluators

Subjects vs Career Age

Theorists–Observers, Men–Women, US+Canada–Europe–Australia–China

	Retired & Active	Senior	Mid-Career	Junior
Very broad	C McKee	R Blandford, JP Ostriker, N Murray, E van Dishoeck, KC Freeman, R Kennicutt, A Fabian	E Quataert	
Solar Ap		J Kuhn		
Planets:SS&Exo		D Jewitt		
Stars	NJ Evans	M Asplund	V Bromm	
Galaxies	SM Faber	LC Ho, J Kormendy	M Cappellari	
Cosmology		D Spergel, B Schmidt		
Interstellar gas and dust		F Combes, B Draine		

I emphasize people who have extensive experience in leading, planning, and judging astronomy research across subject boundaries (e. g., leaders of US Decadal Surveys).

22 LR Evaluators

Subjects vs Career Age

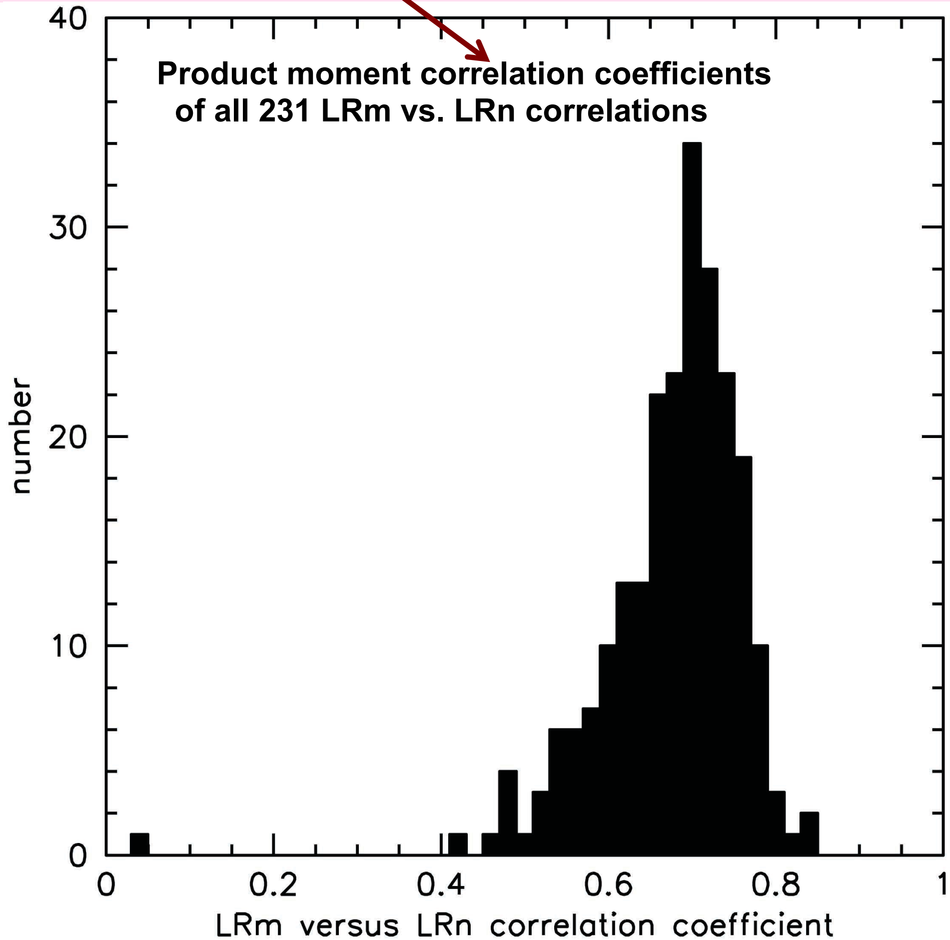
Theorists–Observers, Men–Women, US+Canada–Europe–Australia–China

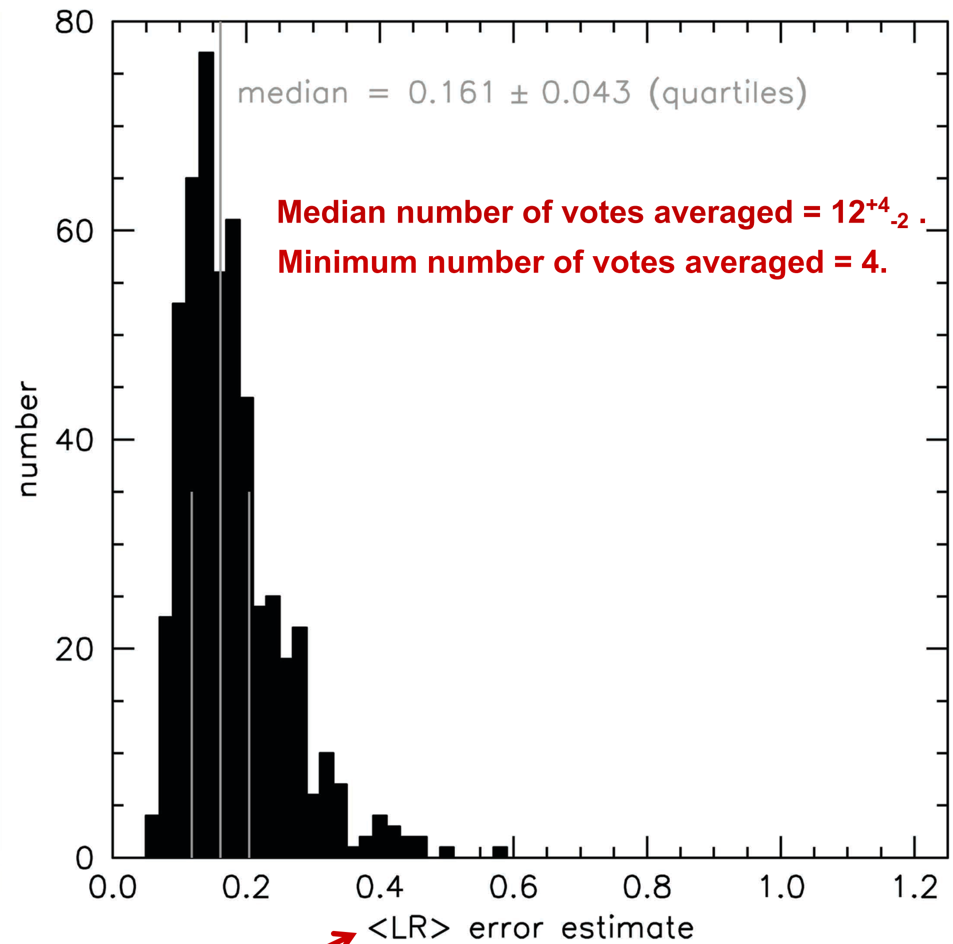
	Retired & Active	Senior	Mid-Career	Junior
Very broad	C McKee	R Blandford, JP Ostriker, N Murray, E van Dishoeck, KC Freeman, R Kennicutt, A Fabian	E Quataert	
Solar Ap		J Kuhn		
Planets:SS&Exo		D Jewitt		
Stars	NJ Evans	M Asplund	V Bromm	
Galaxies	SM Faber	LC Ho, J Kormendy	M Cappellari	
Cosmology		D Spergel, B Schmidt		
Interstellar gas and dust		F Combes, B Draine		

I emphasize people who have had major impact on the histories of their subjects.

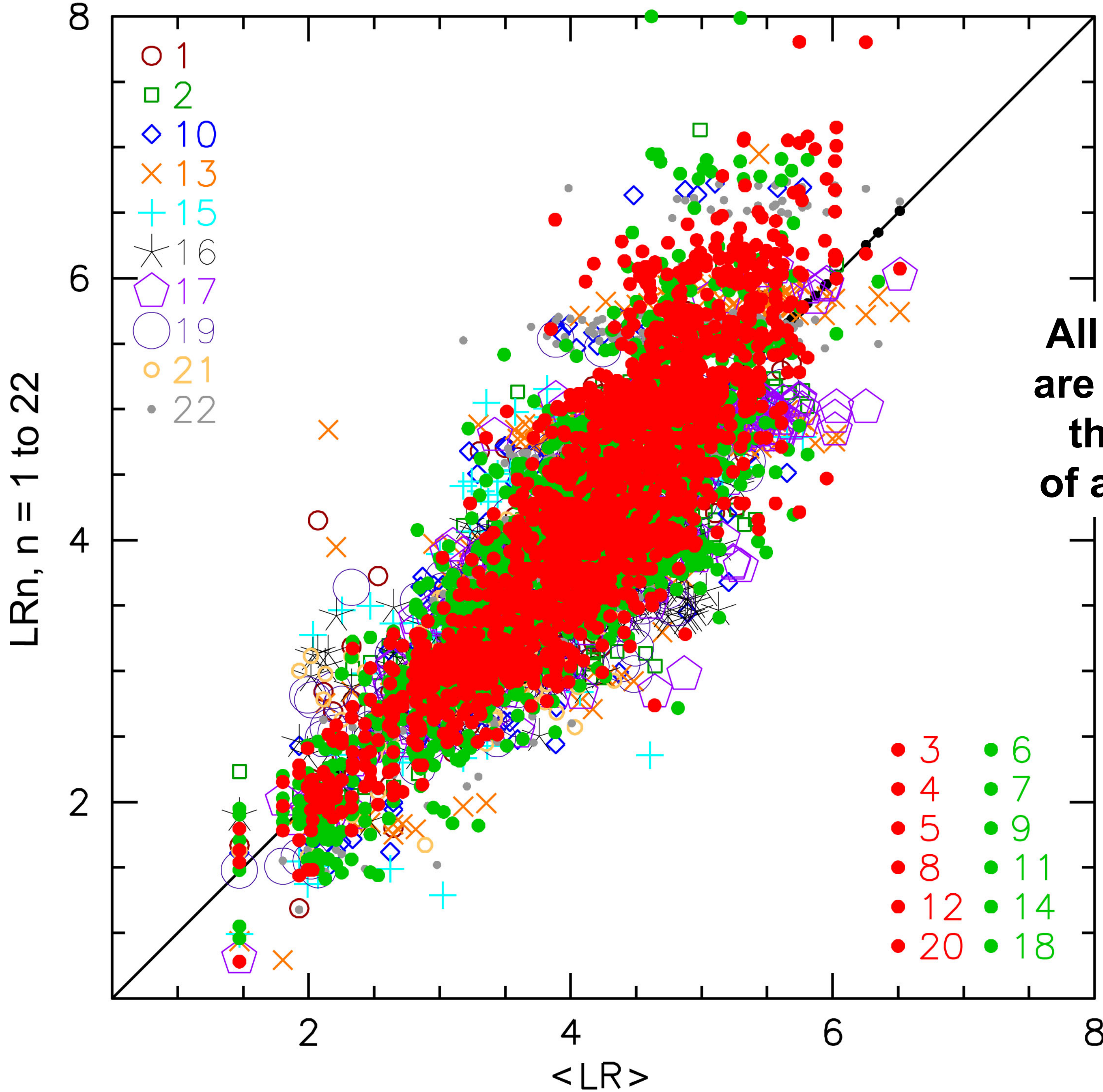
Of the 22 LR voters, 16 are National Academy members.

LR voters measure similar signals with different S/N and dynamic range.





Uncertainty in <LR> measures only voter consistency. No value judgment is implied.



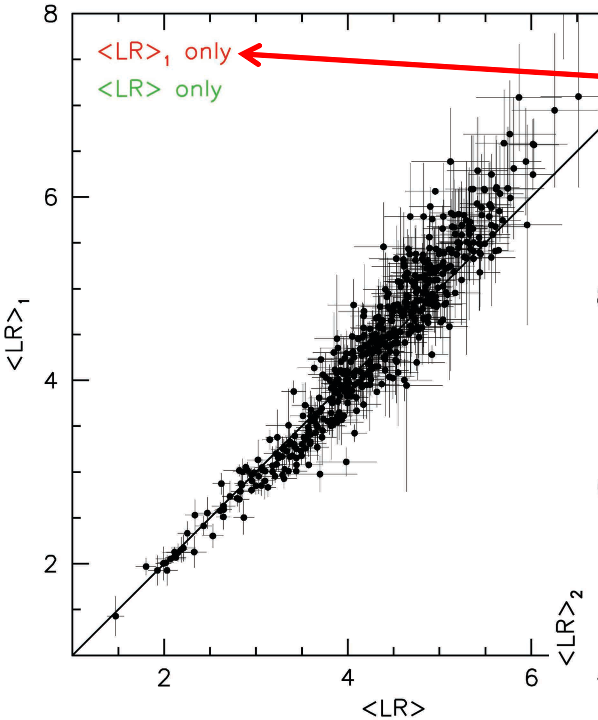
All 6380 LR votes are plotted against the mean <LR> of all 22 vote sets.

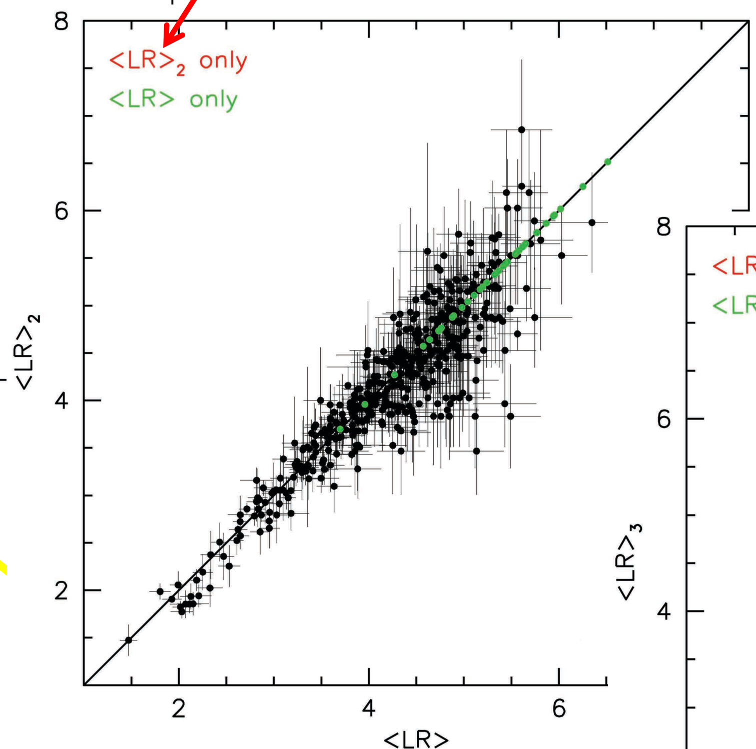
LR vote sets are divided into

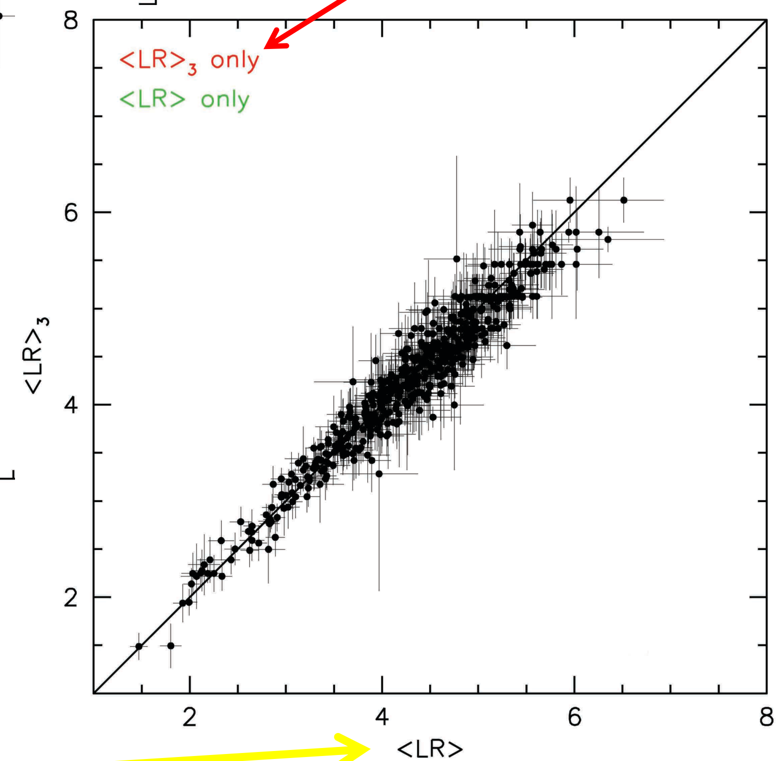
<most-self-consistent 6> = <LR>$_1$,

<medium-self-consistent 6> = <LRL$_2$,

<least-self-consistent 10> = <LR>$_3$.







Three independent vote averages measure similar signals.

Analysis uses unweighted mean LR for all 22 people.

LR voter biases are not a big problem for this work.

Gender bias: 3 women voters judge women researchers to have higher impact than 19 men do … by 0.20 ± 0.05 LR units.

Institutional bias in favor of one's own institution is ~ 1/3 LR unit … but only at intermediate impact. This may partly be a "signal", not a "bias" – people may know their institutional colleagues better than do external voters.

Subject-dependent bias, **geographic bias** and **bias based on age of LR voters** are negligible. **Theorists** and **observers** agree.

I measure the impact that happens, not the impact that should happen.

Job candidates are best served if the machinery includes biases that they will experience.

Averaging over 12^{+4}_{-2} voters reduces any bias in <LR>.

Strategy

I trust that LR measures how history will remember and value research contributions at all career ages. Therefore:

My strategy is to make small "tweaks" to metrics until they correlate as well as possible with LR and can be used as proxies for LR voter opinions.

As careers evolve and people accrue impact, they should evolve upward along LR correlations.

I promise LR voters and study sample researchers that I will keep LR votes anonymous. Therefore:

In all plots, point coordinates are <u>disguised</u> by enough to prevent "reverse engineering" but not so much as to obscure correlations.

All <u>calculations</u> (e. g. correlation fits & RMS) are made with undisguised data.

Application Step 1:

Make an "ADS private library" of all publications for each person who is to be compared.

This is the most work.

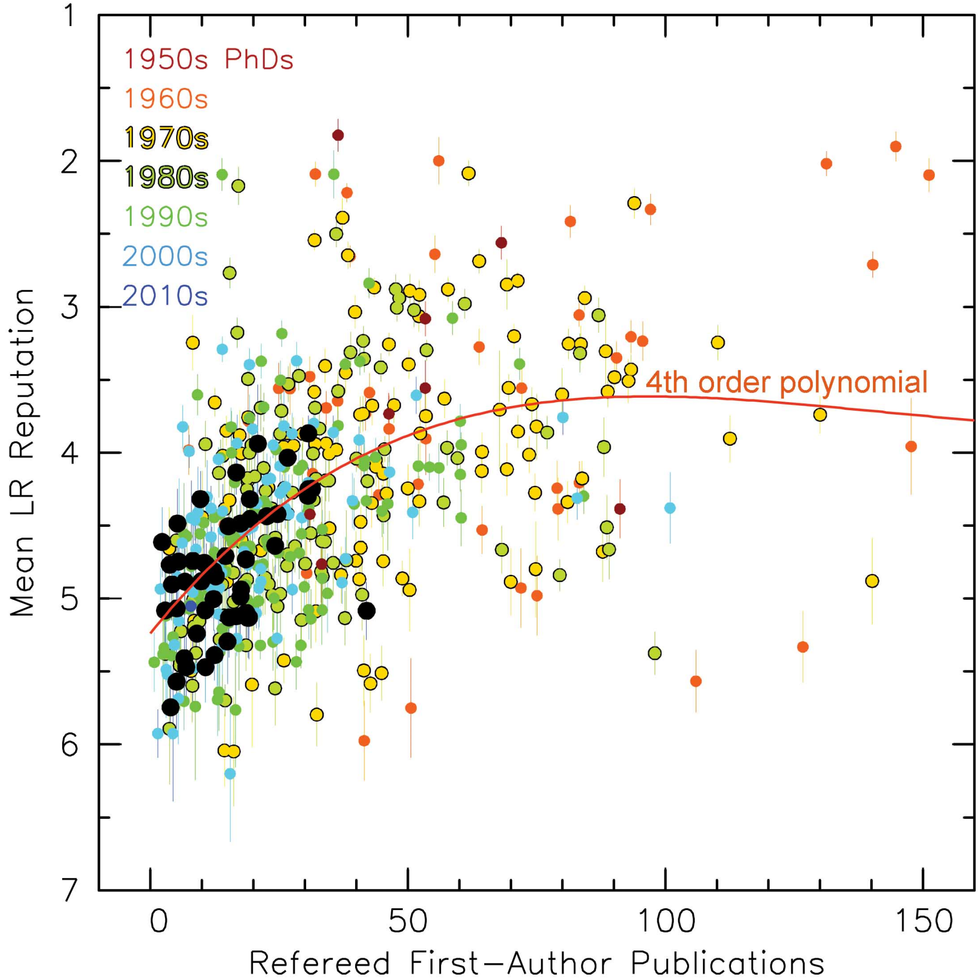
It took me 1 year of full-time work to make private libraries for 510 people, because there are many name duplications in ADS and because I included unrefereed papers. This is <u>more fair</u> but <u>more work</u> than using only refereed papers.

This should be almost no work. In the USA, if we want to use metrics, we should ask candidates to submit private libraries with their applications.

In Hungary, mtmt.hu solves this problem.

Counting papers tells us almost nothing about impact.

It is more important to publish high-impact papers than to publish many papers.



This figure calibrates publication standards for tenure-stream at the present institutions. Standards are high!

Metrics book calibrates 10 metric machines.

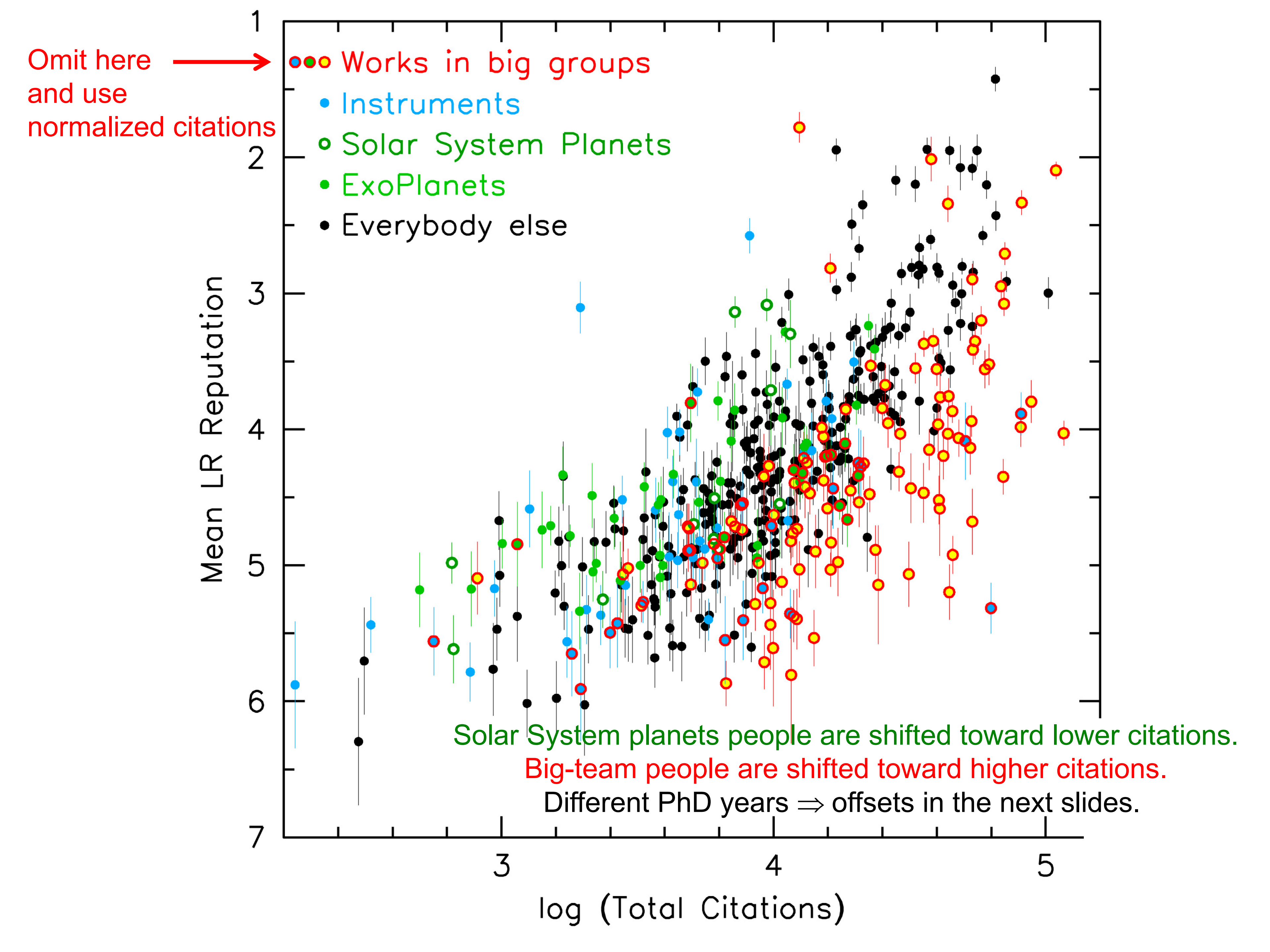
The table lists RMS(LR) for fits of voted LR vs metrics.

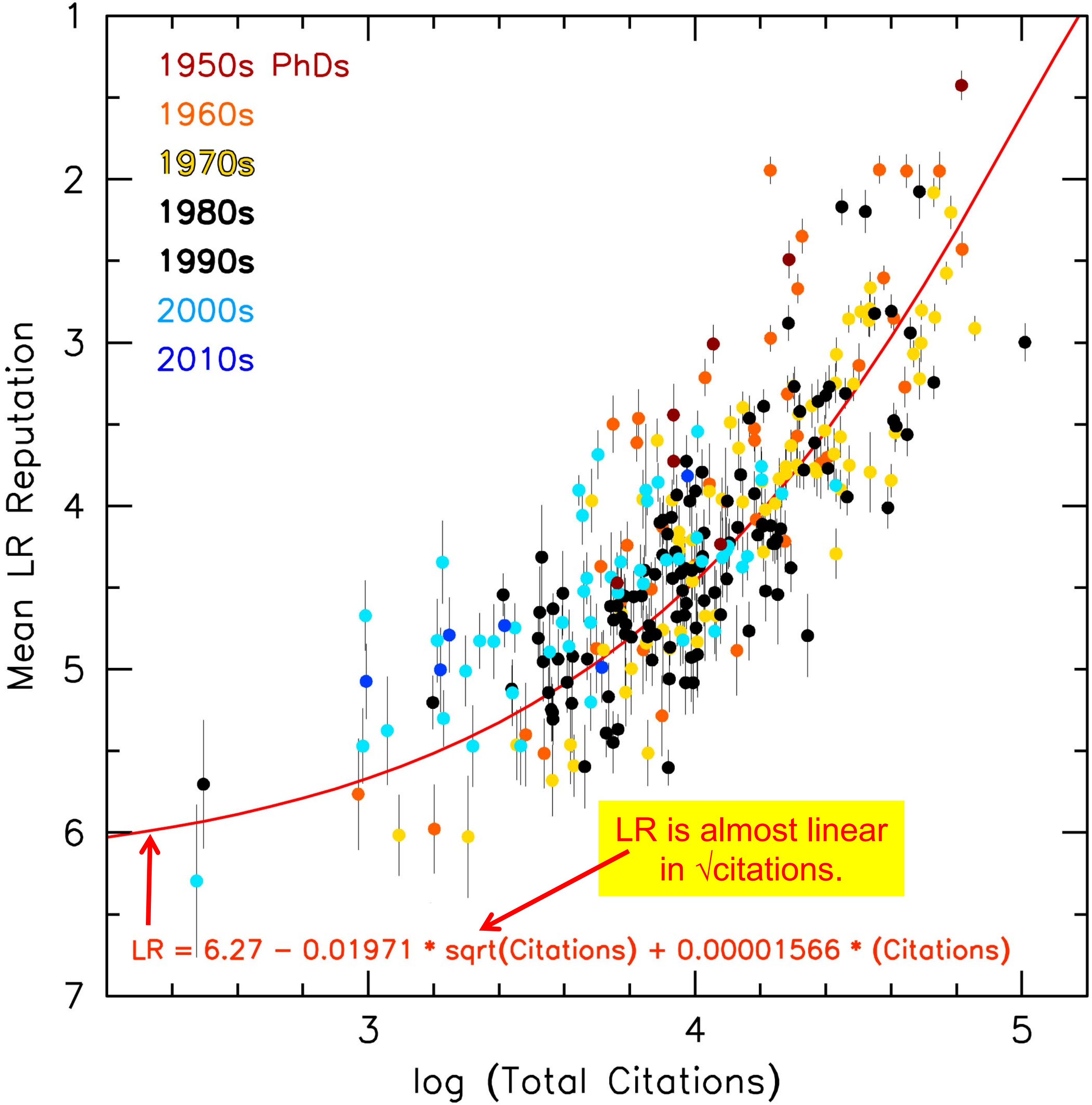
Total citations (i. e., citations of all publications) work best for non-big-team people, including instrumentalists.

Cohort	First-Author Citations 2013 – 2017	Total Citations (sqrt)	Refereed Citations (sqrt)	Total Citations (log)	Refereed Citations (log)	Normalized Citations of All Papers	Tori Index	First-Author Citations of All Papers	I100	Reads of All Papers
Big team	0.61	. . .	. . .	. . .	. . .	0.45	0.45	0.43	. . .	. . .
Instrumentalists	0.76	0.42	0.46	0.50	0.54	. . .	. . .	. . .	0.35	. . .
Everybody else	0.57	0.45	0.44	0.45	0.46	0.42	0.45	0.54	0.44	0.37

Note 1: Two metrics are analyzed in 2 different ways to show that results are insensitive to choice of analysis method.

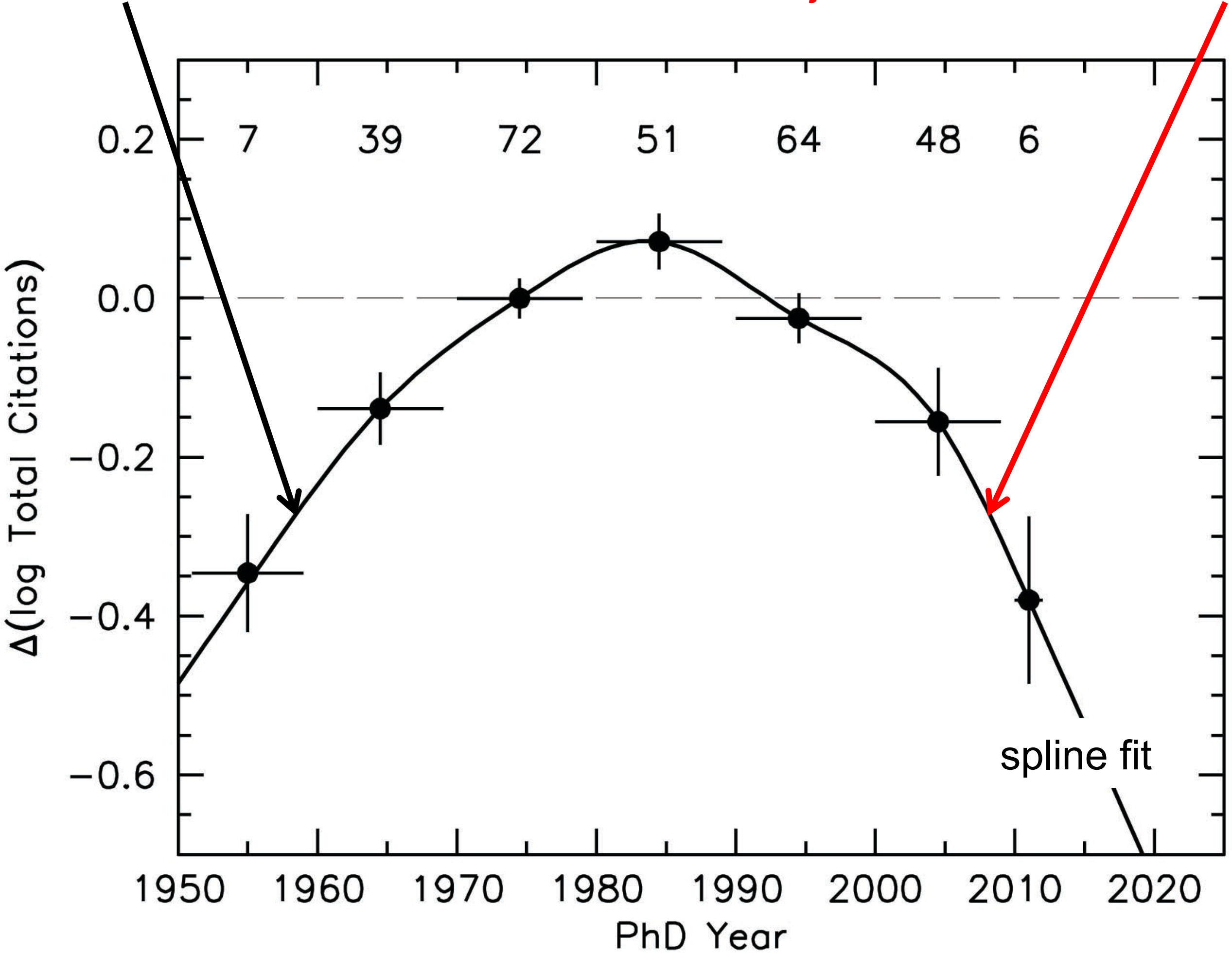
Note 2: Different metrics are most useful for different cohorts of researchers; "…" indicates that this metric is not useful for this cohort.


Omit here
and use
normalized citations
Works in big groups
Instruments
Solar System Planets
ExoPlanets
Everybody else
Mean LR Reputation
1
2
3
4
5
6
7
3
4
5
log (Total Citations)
Solar System planets people are shifted toward lower citations.
Big-team people are shifted toward higher citations.
Different PhD years ⇒ offsets in the next slides.

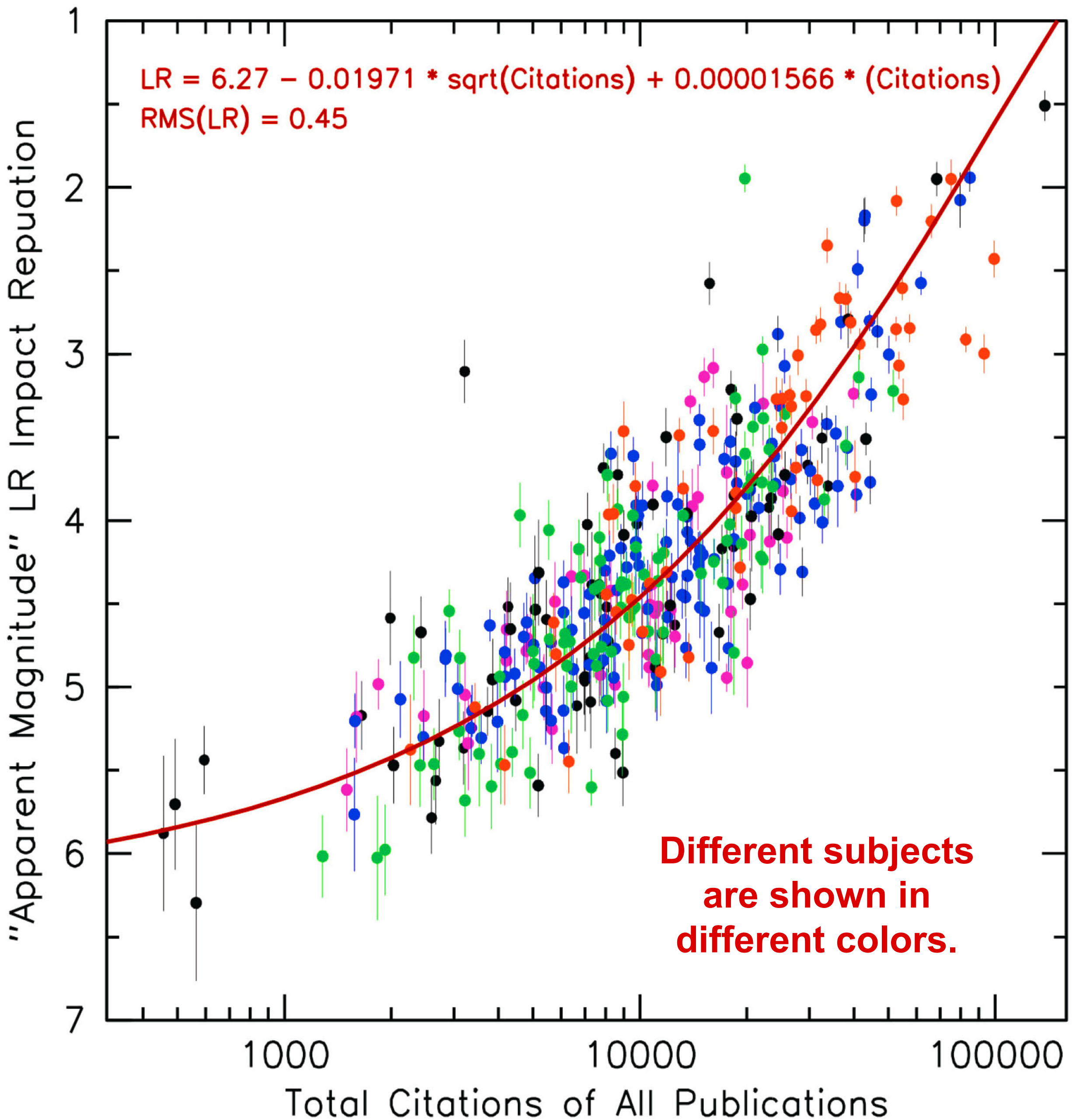

1950s PhDs
1960s
1970s
1980s
1990s
2000s
2010s
Mean LR Reputation
1
2
3
4
5
6
7
3
4
5
log (Total Citations)
LR is almost linear in √citations.
LR = 6.27 − 0.01971 * sqrt(Citations) + 0.00001566 * (Citations)

Oldest PhDs got fewer citations per unit impact
because there were fewer astronomers,
fewer journals and journal pages,
and fewer citations.

Youngest PhDs get fewer citations
per unit impact: I suspect that
they are judged partly via perceived potential
and not just via contributions already made.



Result: Impact correlates well with citation counts



Here:

shifts have been applied to calibrate out career age

and

subject-dependent shifts have been applied for a few subjects.

This is for people who do not work mainly in "big teams" of > 30 people.

To use total citations to predict LR: Adjust citation numbers for PhD year via Fig. 63 and Appendix 3. Add Δlog(total citations) from Table 9. Then apply Equation (5) to estimate LR as in Figure 66. LR uncertainties and Equation (5) are given in the key.

TABLE 9: SUBJECT-DEPENDENT SHIFTS IN LOG(TOTAL CITATIONS)

Subject	Δ log(citations)	Subject	Δ log(citations)
Sun	+0.11	ISM	+0.09
Solar System	+0.30	Instruments	+0.19
Exoplanets	+0.16	Other subjects	+0.00
Cosmology	+0.18	Stars and galaxies	insignificant shifts < 0

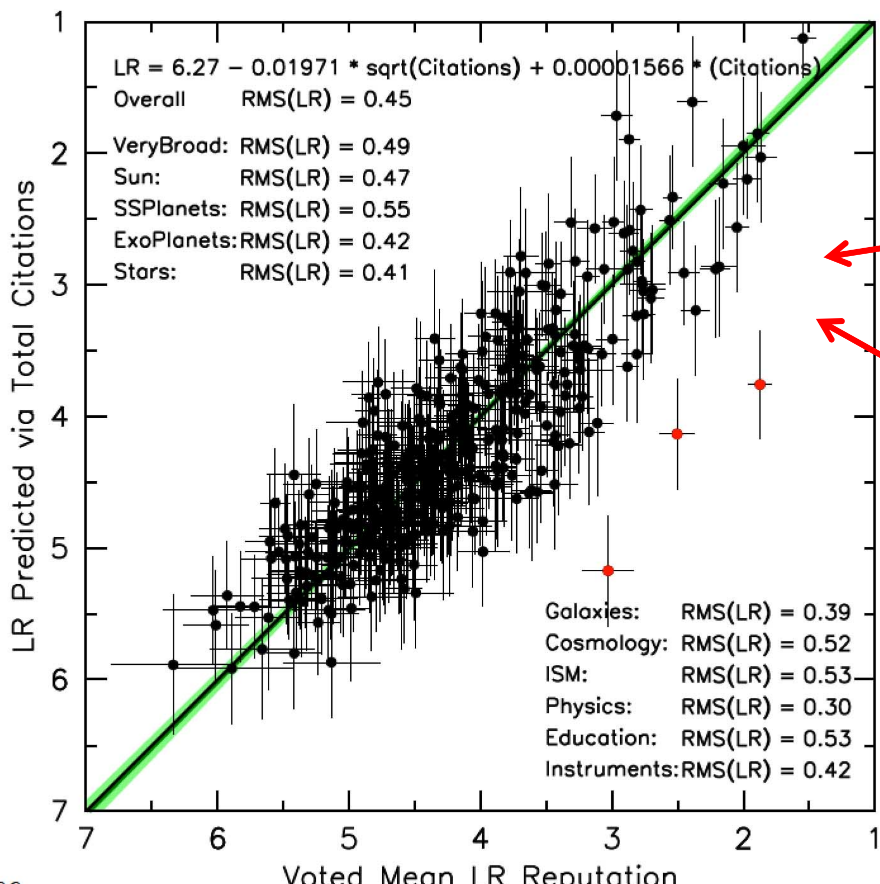


Figure 66

Correlation of voted LR with the prediction (*box*) based on total citations after shifts for PhD year and subject. RMS values for different subjects (*keys*) are used as estimates of the uncertainties in predicted LR. Three people who deviate by > 2.5 σ (*red points*) are omitted from RMS estimates.

For each metric machine, I provide a box that summarizes use and a calibration of how well that metric acts as a proxy for the LR voters.

For total citations and for citations of referred papers, RMS(LR) is already good.

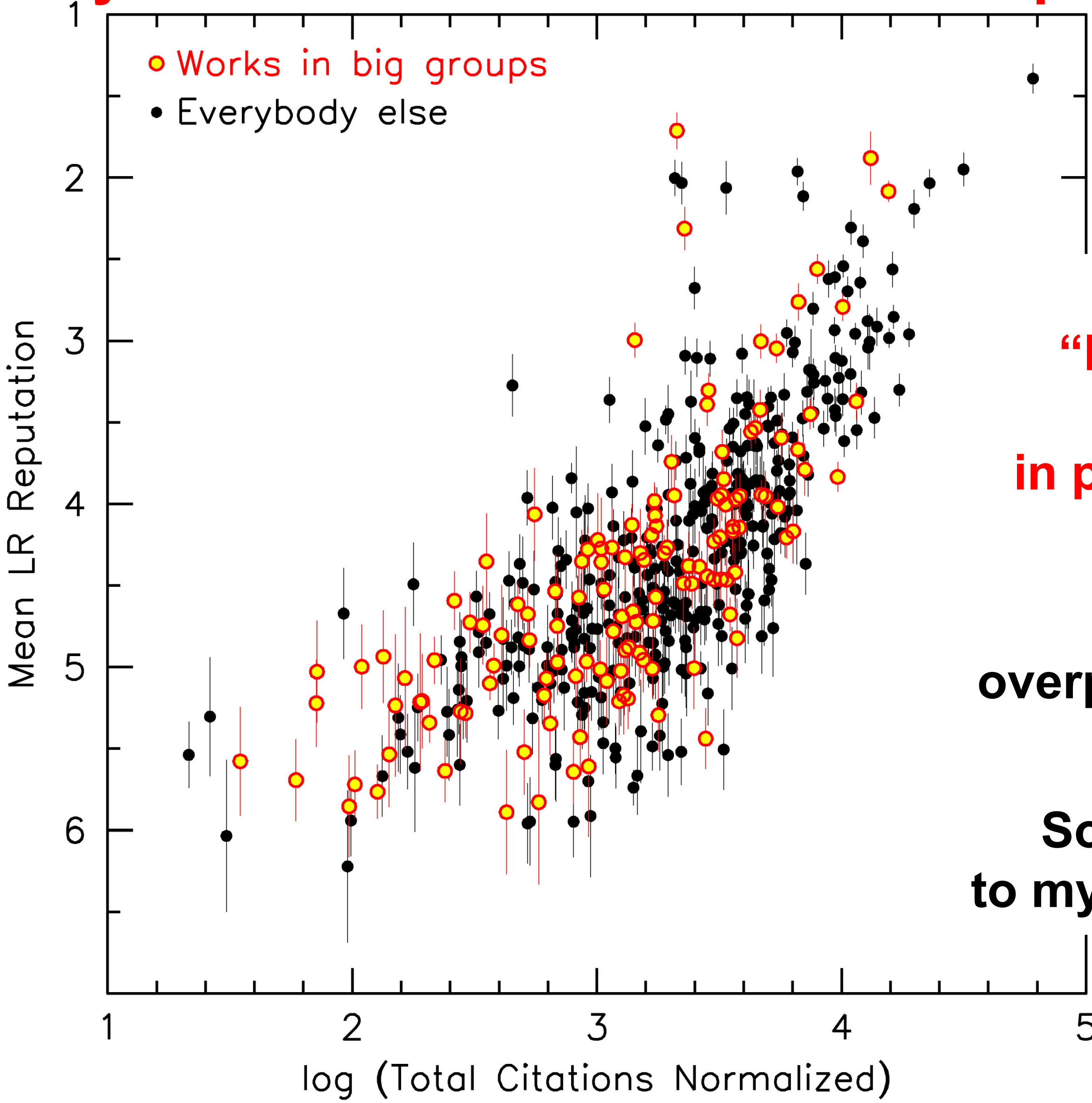
Metrics book calibrates 10 metric machines.

The table lists RMS(LR) for fits of voted LR vs metrics.

Cohort	First-Author Citations 2013 – 2017	Total Citations (sqrt)	Refereed Citations (sqrt)	Total Citations (log)	Refereed Citations (log)	Normalized Citations of All Papers	Tori Index	First-Author Citations of All Papers	I100	Reads of All Papers
Big team	0.61	...	...	...	...	0.45	0.45	0.43	...	...
Instrumentalists	0.76	0.42	0.46	0.50	0.54	...	...	...	0.35	...
Everybody else	0.57	0.45	0.44	0.45	0.46	0.42	0.45	0.54	0.44	0.37

Normalized citations of all publications
(i. e., citations for each paper are divided by the number of co-authors)
work best for big-team people
and are also good for non-big-team people
(except instrumentalists).

For big-team people, we need to used citations normalized by the number of authors on each paper.



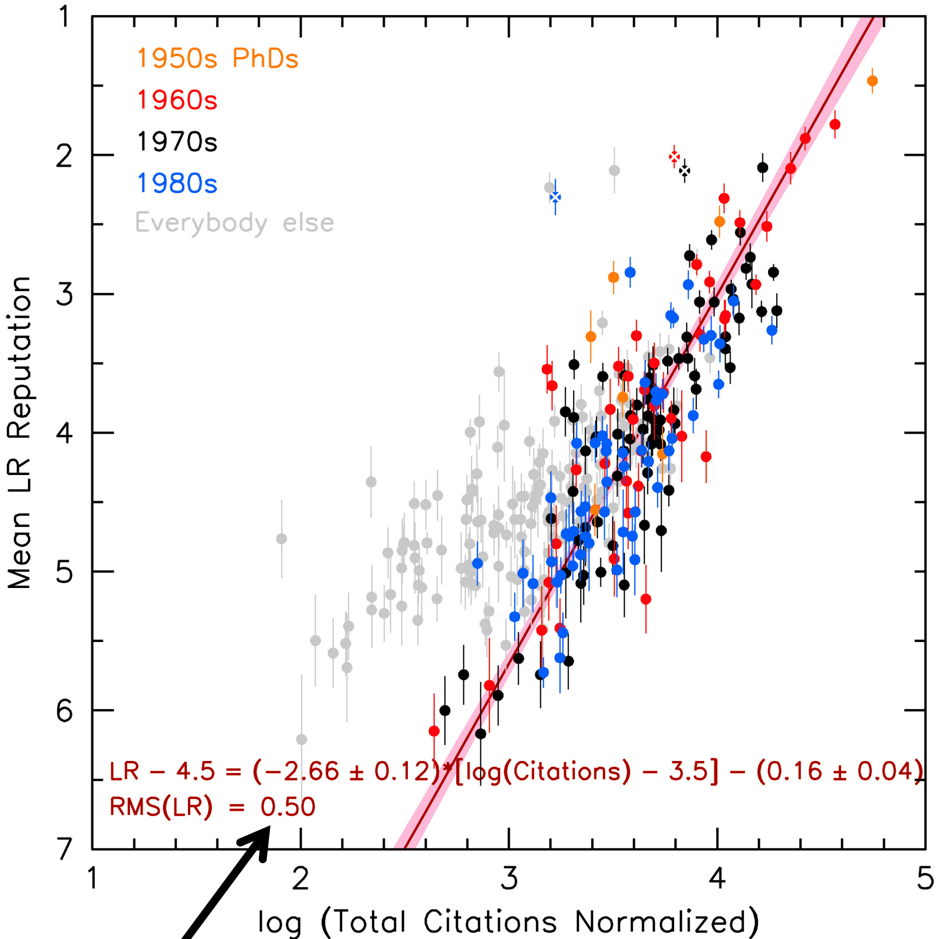
Definition:
"Big-team people" have most of their impact in papers with ≥ 30 authors.

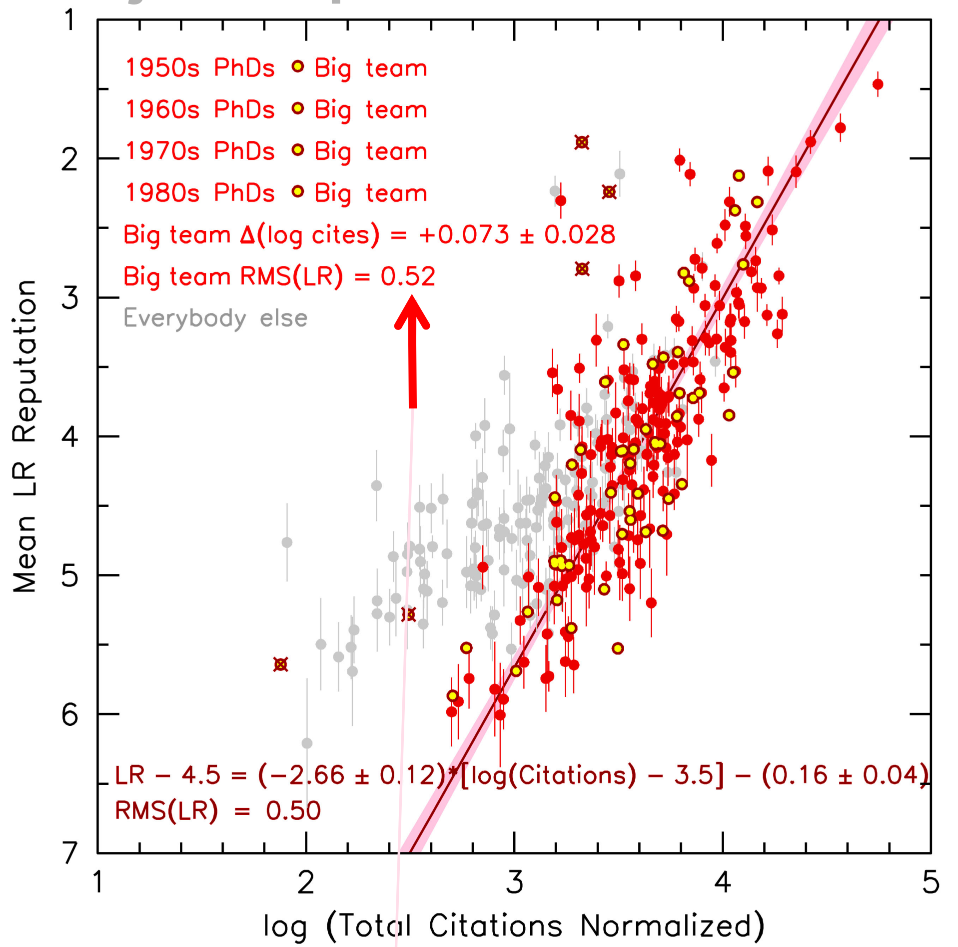
Their distribution of normalized citations overplots non-big-team people very well.

So results are insensitive to my definition of "big teams".

People with 1950s – 1980s PhDs have the same correlation of LR with normalized citations. Make least-squares fit (red) to non-big-team people.

Subsample is too small for subject-dependent shifts.

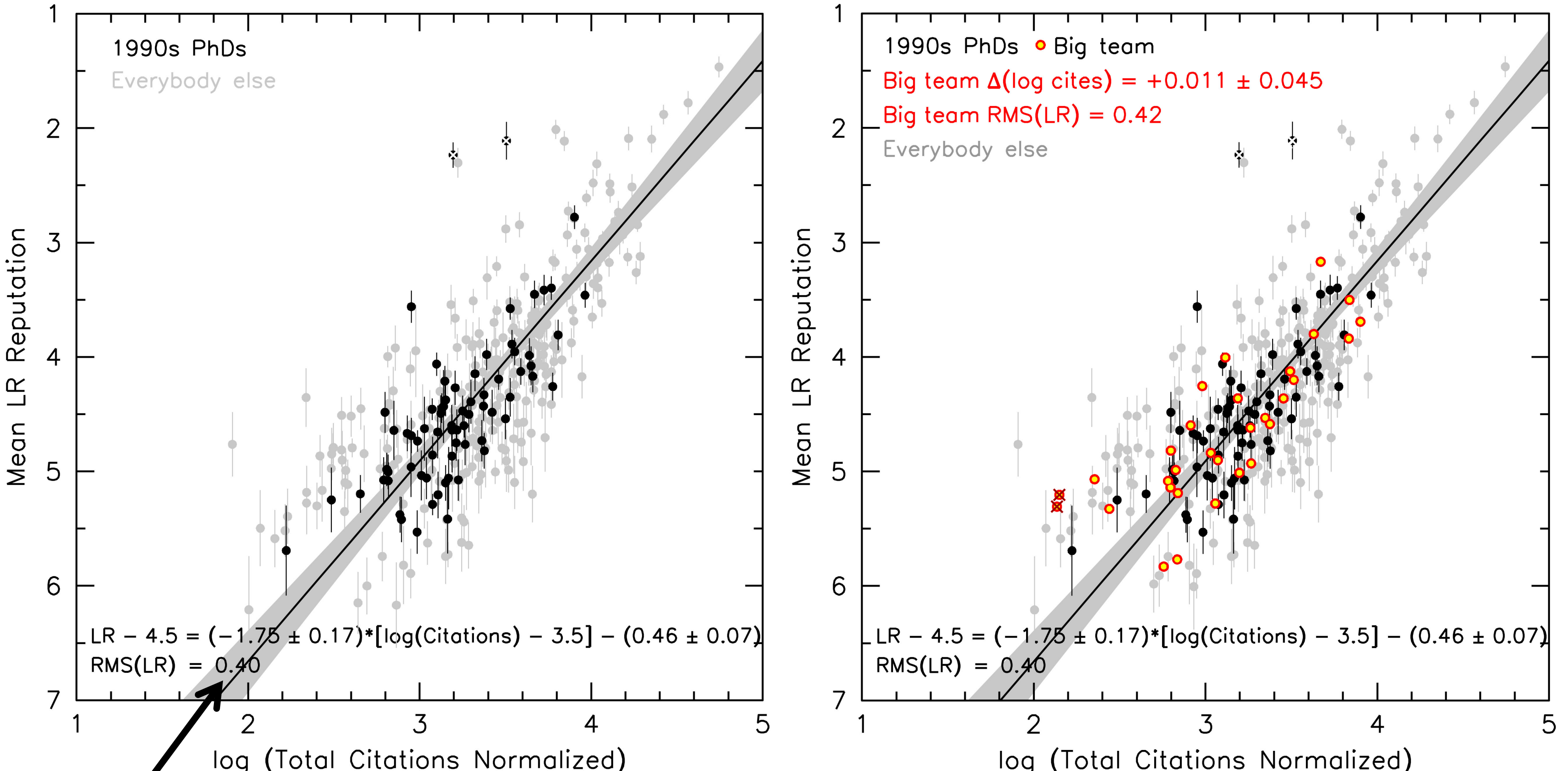




RMS(LR) = 0.50 for non-big-team people (omit 3).

RMS(LR) = 0.52 is almost the same for big-team people (omit 5).

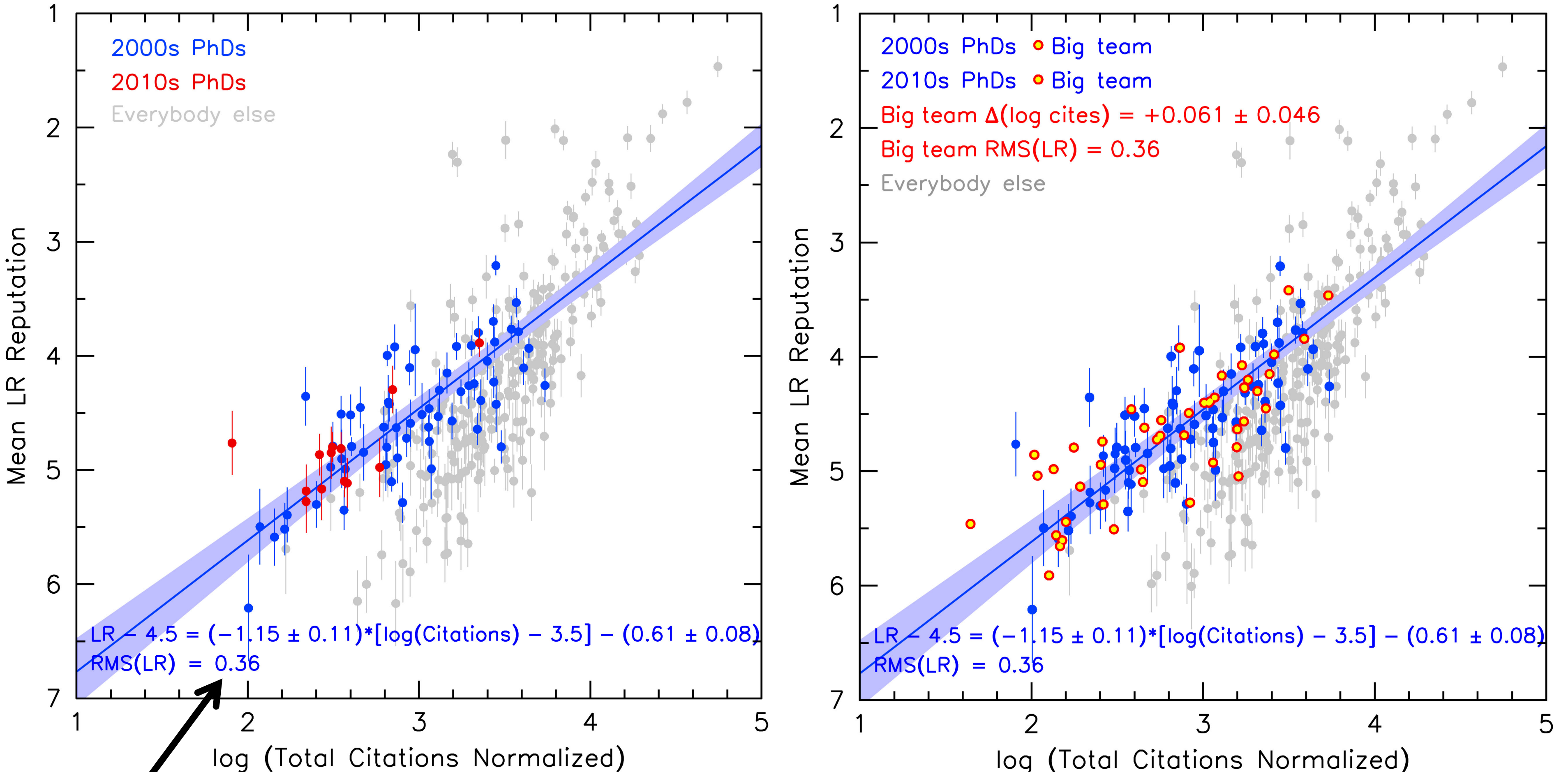
For 1900s PhDs, citation behavior starts to change.



RMS(LR) = 0.40 for non-big-team people (omit 2).

Correlation and RMS(LR) = 0.42 is almost the same for big-team people.

For 2000s and 2010s PhDs, a least-squares fit (blue) to non-big-team people is still shallower than for 1950s – 1990s PhDs.



RMS(LR) = 0.36 for non-big-team people.

Correlation and RMS(LR) = 0.36 is the same for big-team people.

Normalized citations of all papers ("cites") via the least-squares fits in Figures 93 – 95:

For 1950s – 1980s PhDs:

$$\mathrm{LR} - 4.5 = (-2.66 \pm 0.12)\,[\log(\mathrm{cites}) + \Delta - 3.5] - (0.16 \pm 0.04); \quad \mathrm{RMS(LR)} = 0.50 \text{ or } 0.52, \quad (23)$$

with $\Delta = 0.073 \pm 0.028$ for big-team people and zero for everybody else.

For 1990s PhDs:

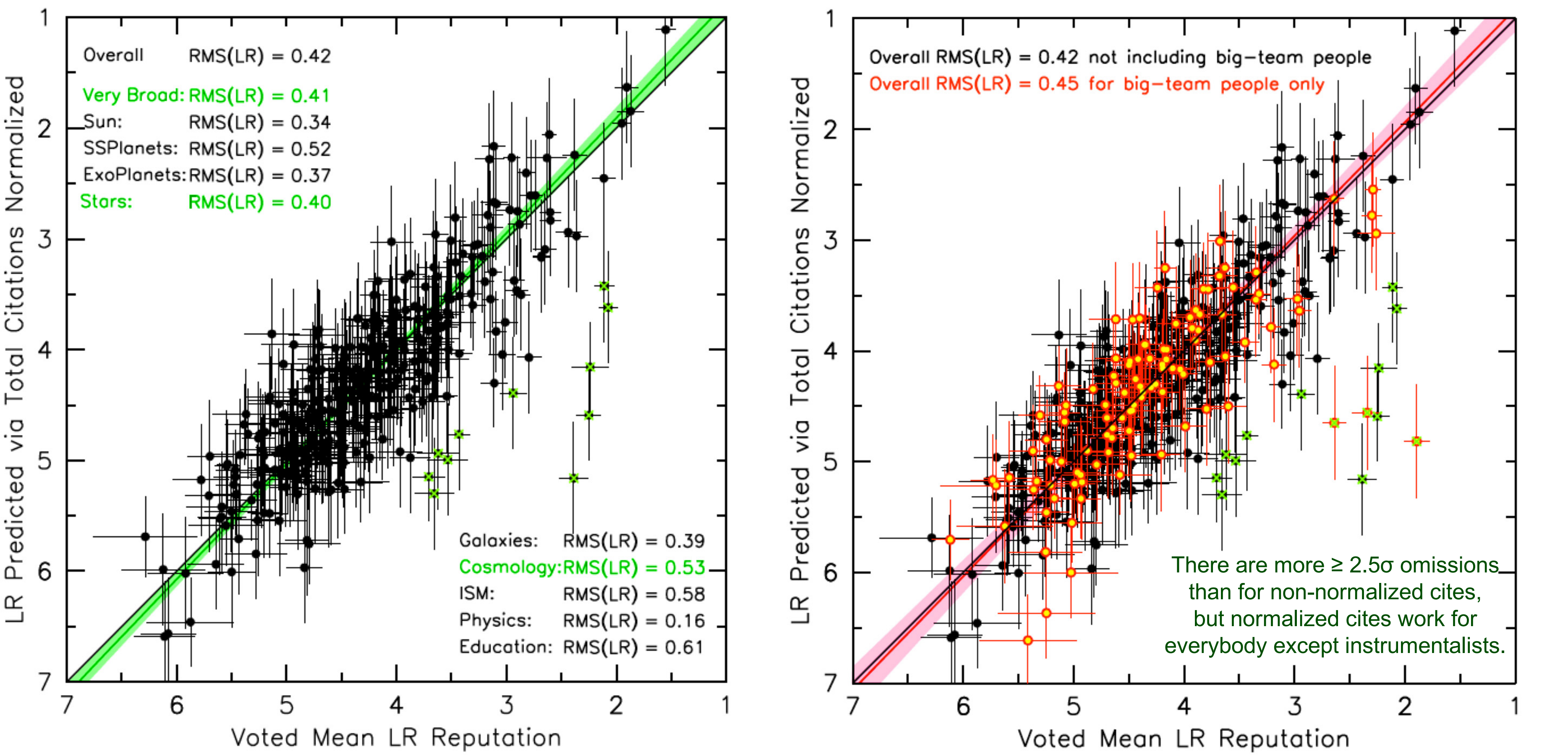
$$\mathrm{LR} - 4.5 = (-1.75 \pm 0.17)\,[\log(\mathrm{cites}) \qquad - 3.5] - (0.46 \pm 0.07). \quad \mathrm{RMS(LR)} = 0.40 \text{ or } 0.42. \quad (24)$$

For 2000ss – 2010s PhDs:

$$\mathrm{LR} - 4.5 = (-1.15 \pm 0.11)\,[\log(\mathrm{cites}) + \Delta - 3.5] - (0.61 \pm 0.08); \quad \mathrm{RMS(LR)} = 0.36 \text{ or } 0.36, \quad (25)$$

with $\Delta = 0.061 \pm 0.046$ for big-team people and zero for everybody else. In each equation, the first RMS(LR) value applies to non-big-team people and the second applies to big-team people.

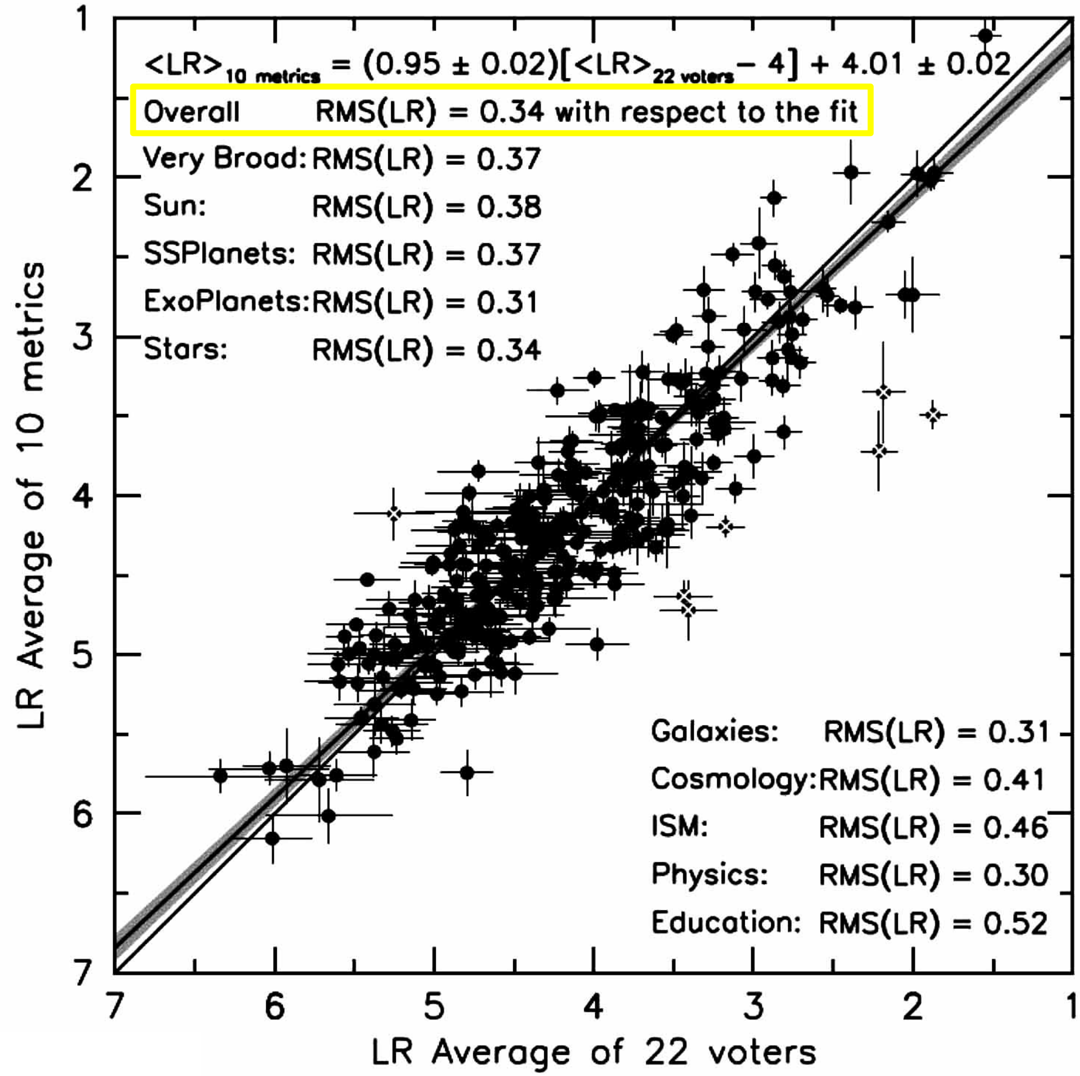
Small Δ offsets
for big-team people
are barely significant.



Average several metrics ⇒ more accurate proxy:

<10 metric machines> ⇒ voted LR with RMS = 0.34.

This is better than I dared to expect.



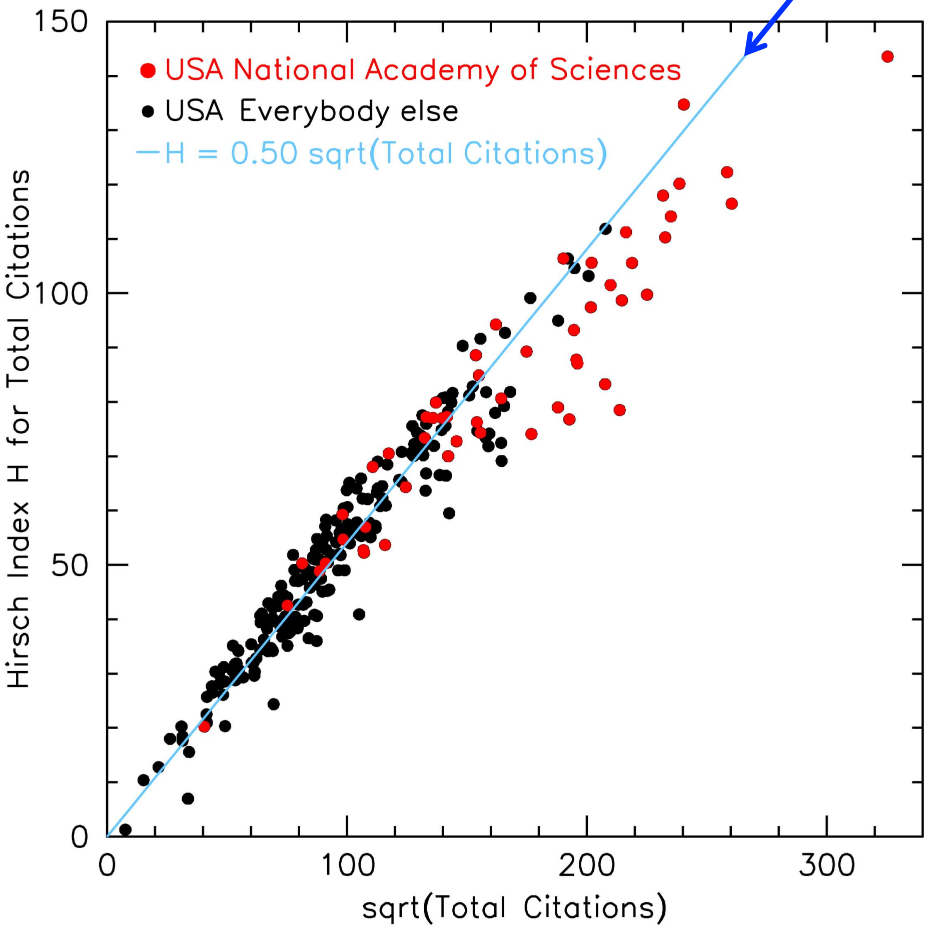
I provide combinations of 2 – 5 metrics optimized for various cohorts.

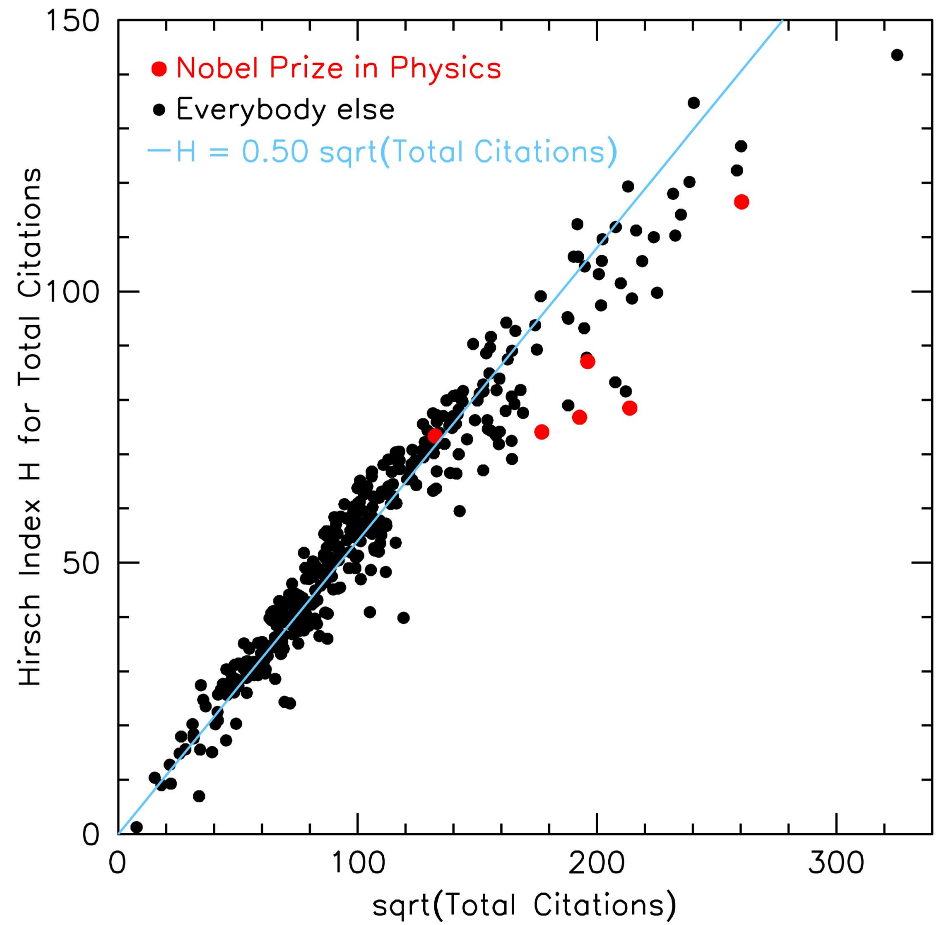
Hirsch [2005, Proc. Nat. Acad. Sci. 102(46), 16559] **Index**

H = largest number such that you have H papers with ≧ H citations.

It is "doubly hard" to grow H: standards get harder as H increases.
No surprise: H ∝ $\sqrt{\text{citations}}$. This can be derived from combinatorics [Yong 2014, Notices Amer. Math. Soc. 61(11), 1040].

This is not a fit.

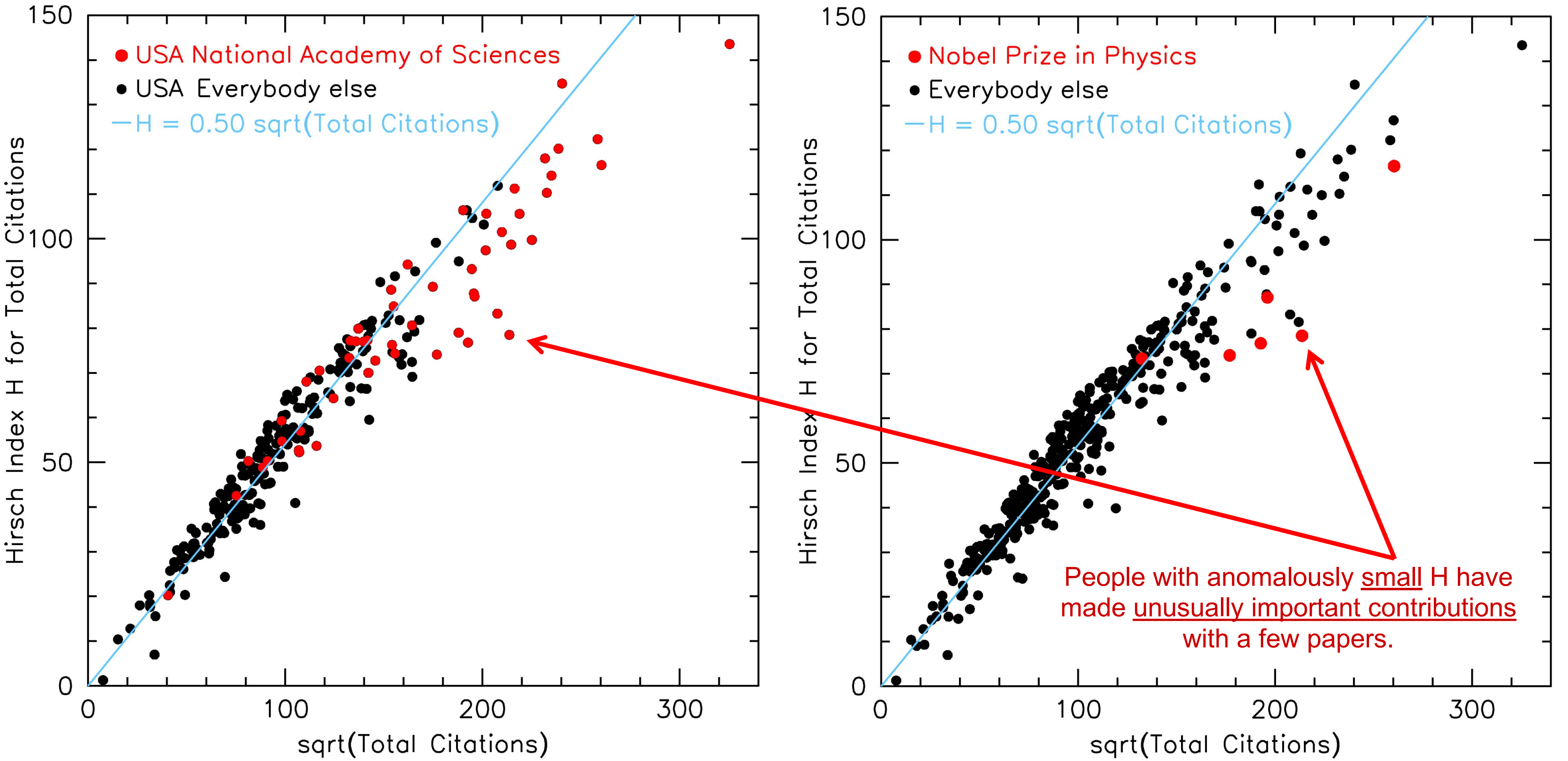




Conclusion

H is almost always used in the opposite to the most useful way:
People ask, "How big is this person's H?"
This tells us nothing that total citations have not already told us.

Should ask: "How small is H compared to the prediction of the correlation?



How to use metrics to rank candidates:

1 – Make ADS private libraries of publications for everybody (in astronomy) and/or use mtmt.
Get the metrics that you want to use.
More metrics give more robust results.

2a – If you don't derive LR:
Rank people separately using each metric.
Consistency (or not) of rankings $\Rightarrow$ uncertainty.

Or:

2b – Derive LR $\pm$ RMS(LR) for each metric.
Then you can average different metrics $\Rightarrow$ $\langle LR \rangle \pm$ smaller RMS($\langle LR \rangle$). Advantages:

Advantages of using <LR> to rank candidates:

1 – Each metric machine is calibrated to the same LR scale,
so metrics can be averaged to improve accuracy.
It is safe to use different metrics for different cohorts.

2 – Deriving <LR> ± RMS(<LR>) ⇒ which rankings are significant.

3 – Subject-dependent tweaks are part of calibration,
so people in different subjects can be compared more accurately.

4 – I provide a comparison sample of 510 people across all subjects.
LR provides context for interpretation on an absolute impact scale.

5 – In this age of scrutinized oversight and accountability, it helps
to record quantitative evidence on which decisions are based.

6 – Automation can reduce the work needed to distill 10^2 applicants down
to the ~ 10 – 20 people who are considered in detail → short list.

These points are especially important for senior hires.

CAUTION

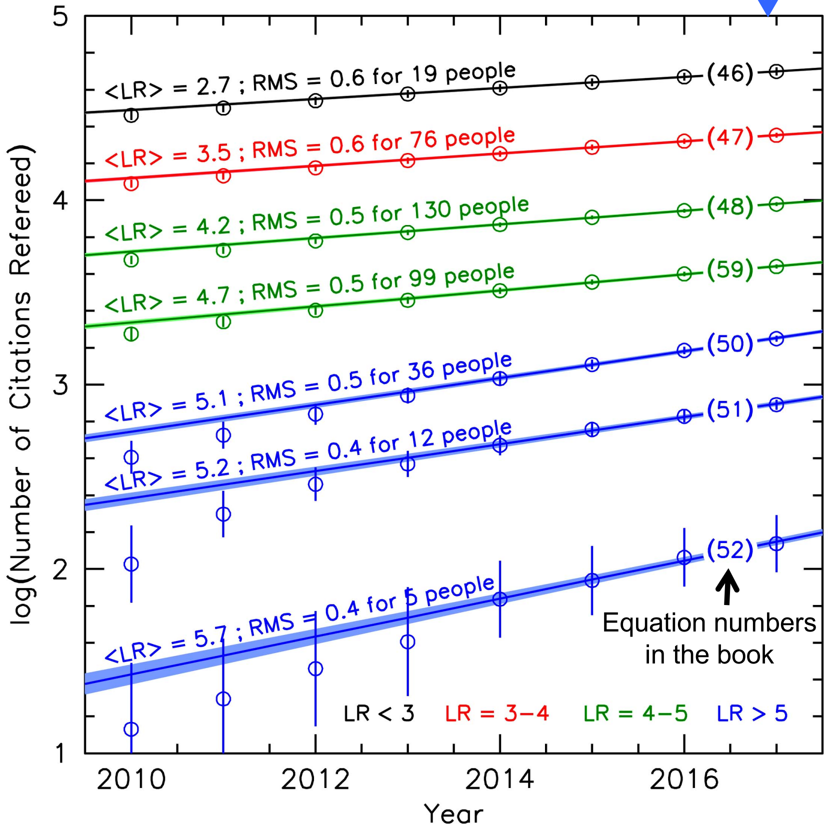
Reputations grow slowly with time.

But log(metrics) increase linearly with time.

If you use 2021 metrics in 2017 machinery without correcting back to 2017 metrics, then you substantially overestimate impact.

Chapter 11 provides machinery needed to put people on the LR scale of the book.

Renormalization is not needed if you only want to order candidates on a relative LR scale as defined by year N metrics, where N > 2017.



I showed that
current metrics measure current impact.

Can current metrics predict future impact?

The answer is “yes – usefully well”. See:

Chapter 13 calibrates prediction from citations of refereed papers;
from normalized citations; and
from first-author citations,

all from 15, 12, and 10 years after the PhD to later = 2017 LR.

and

Kormendy (2021, Proc. Nat. Acad. Sci., resubmitted after refereeing)
averages the above 3 prediction machines.

Metrics of research impact in astronomy: Predicting later impact from metrics measured 10-15 years after the PhD

John Kormendy (2021, Proc. Nat. Acad. Sci., resubmitted after refereeing)

Significance Statement

Astronomers are trained to do scientific research with rigor and precision, using well-known, agreed-upon techniques that yield results with quantitative measures of uncertainty. In contrast, decisions on hiring and career advancement are made using qualitative indicators and uncertain personal opinion. As scientists, we should aim to do better. The book measures career impact. This paper develops machinery to make quantitative predictions of future scientific impact from metrics measured immediately after the ramp-up period that follows the PhD. The aim is to resolve some of the uncertainty in using metrics for one aspect only of career decisions – judging scientific impact.

Thanks

– to the LR voters for their work & for entrusting me with their opinions;

– to the ADS folks – especially Edwin Henneken – for python programs to collect ADS metrics and for the unique service that ADS provides;

– to Ralf Bender for advice, least-squares fitting software, and support of my visits to Munich where much of this work was done;

– to Robert Lupton and Patricia Monger for creating the SM world in which I spent much of the past 5 years;

– to Ralf Bender, Françoise Combes, Sandy Faber, Luis Ho, & Avi Loeb for writing endorsements; also to Avi Loeb for writing the Preface;

– to Joe Jensen and the monographs team at ASP for enthusiastic support of publication; also to Neta Bahcall, my PNAS editor;

– to many people for discussions that helped my work; and

– especially to Mary Kormendy for love, support, and patience!

Longitudinal Studies:
Mean Histories of Citation Rates

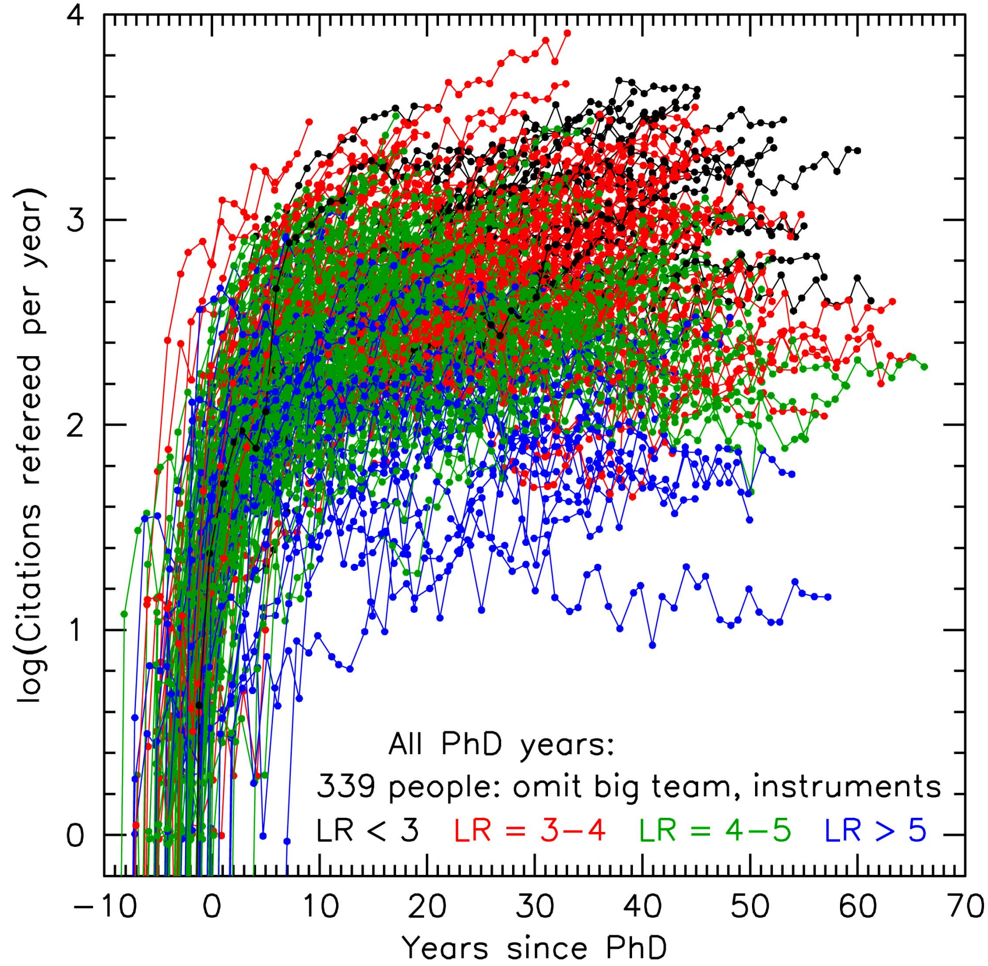
Chapter 12 ⟶ citation rate histories for various LR ranges and PhD years.

Conclusions:

Citation rates ramp up for ~ 10 yr after the PhD.

They "plateau" at higher rates for higher-impact LR.

"Plateaus" are not flat:
highest-impact people increase and
lower-impact people stay constant or decrease slowly in impact rate as time passes.



Metrics of research impact in astronomy: Predicting later impact from metrics measured 10-15 years after the PhD

John Kormendy[a,b] (2021, Proc. Nat. Acad. Sci., resubmitted after refereeing)

My inaugural paper in PNAS averages 3 prediction machines from Chapter 13 of the book:

For non-big-team people:
<citations refereed, normalized citations, first-author citations>

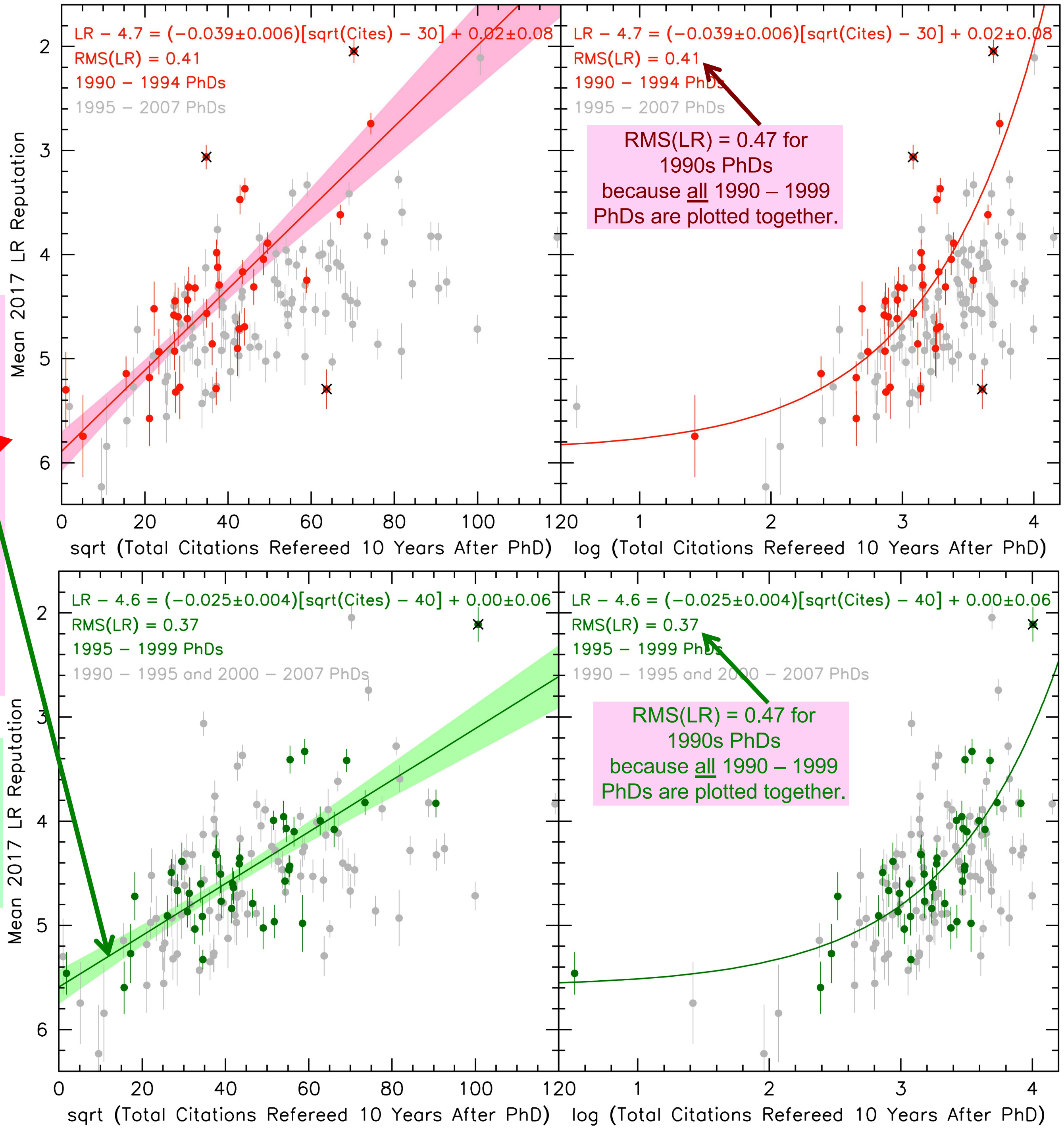
For big-team people:
<normalized citations, first-author citations>

Note: The machinery is calibrated specifically so that different metrics can safely be used for different cohorts of people.

Using Citations 10 yr Post-PhD to Predict Future Impact

For 1990 – 1994 PhDs, predicting 2017 LR is predicting from 2000 – 2004 to 2017.
Happily: RMS(LR) = 0.41, 0.37 is smaller than RMS(LR)=0.47 for contemporary correlations of 2017 metrics with 2017 LR because 1990 – 1999 cohort there is divided into 2 here.

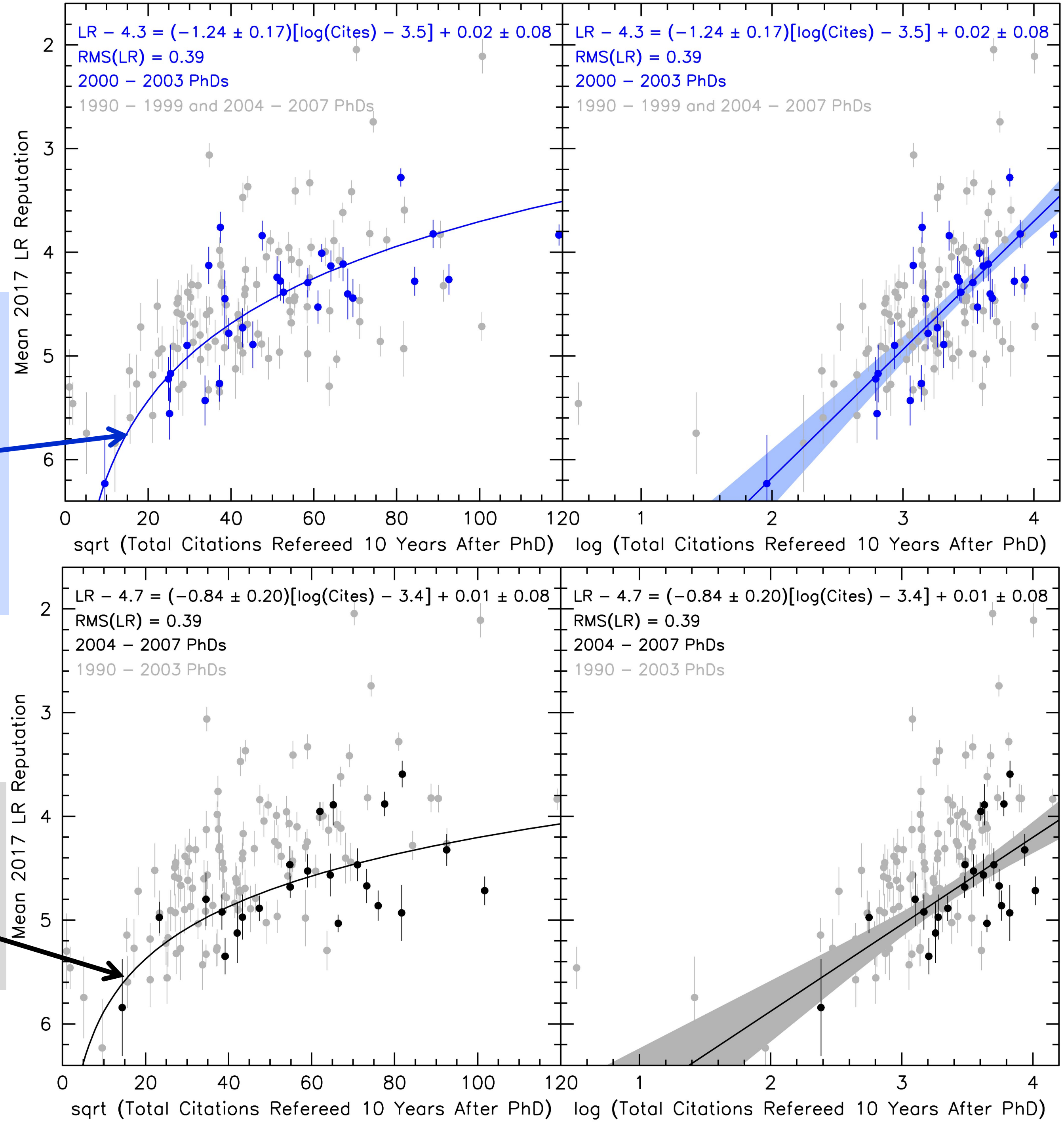
For 1995 – 1999 PhDs, predicting 2017 LR is predicting from 2005 – 2009 to 2017.



Using Citations 10 yr Post-PhD to Predict Future Impact

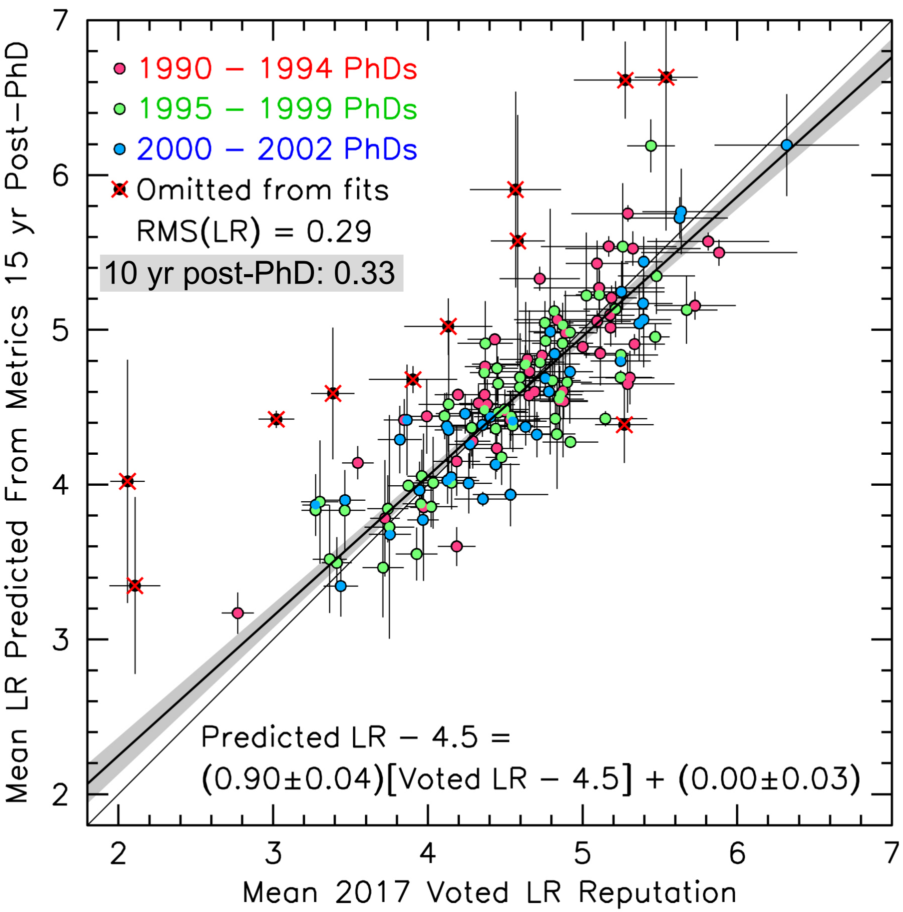
For 2000 – 2003 PhDs, predicting 2017 LR is predicting from 2010 – 2013 to 2017. RMS(LR) = 0.39 is about the same as RMS(LR)=0.38 for contemporary correlations of 2017 metrics with 2017 LR.

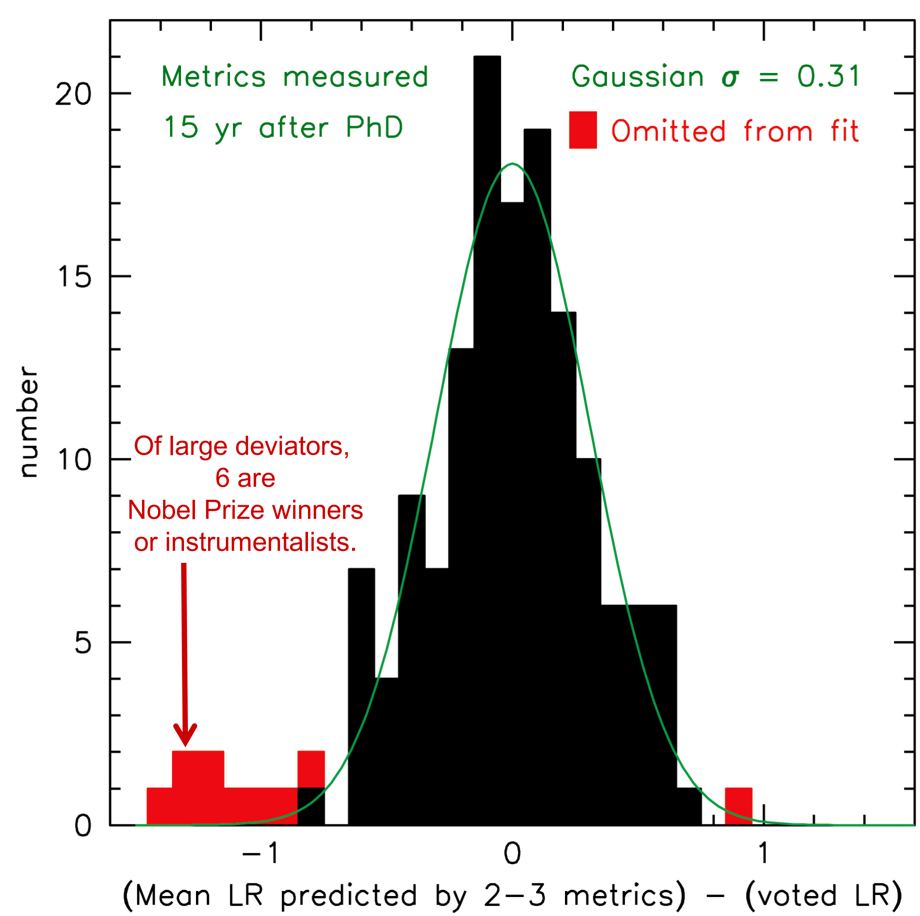
For 2004 – 2007 PhDs, predicting 2017 LR is predicting from 2014 – 2017 to 2017. This is hardly a “prediction”.



Metrics of research impact in astronomy: Predicting later impact from metrics measured 10-15 years after the PhD

John Kormendy[a,b] **(2021, Proc. Nat. Acad. Sci., resubmitted after refereeing)**

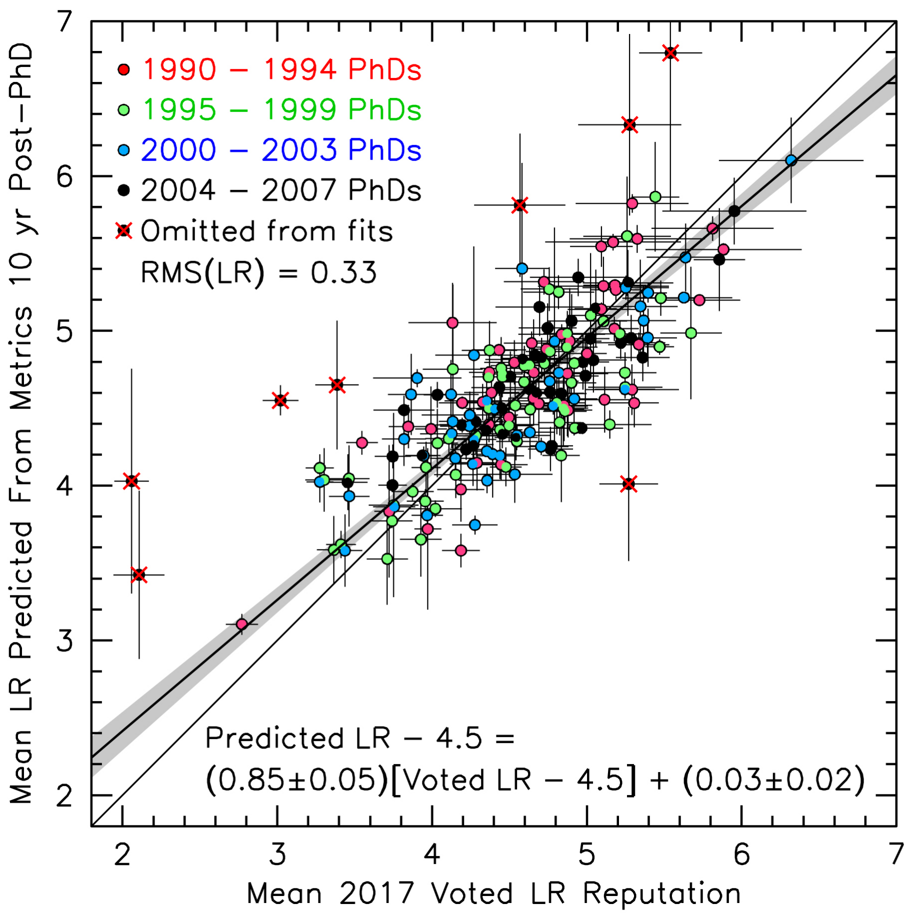


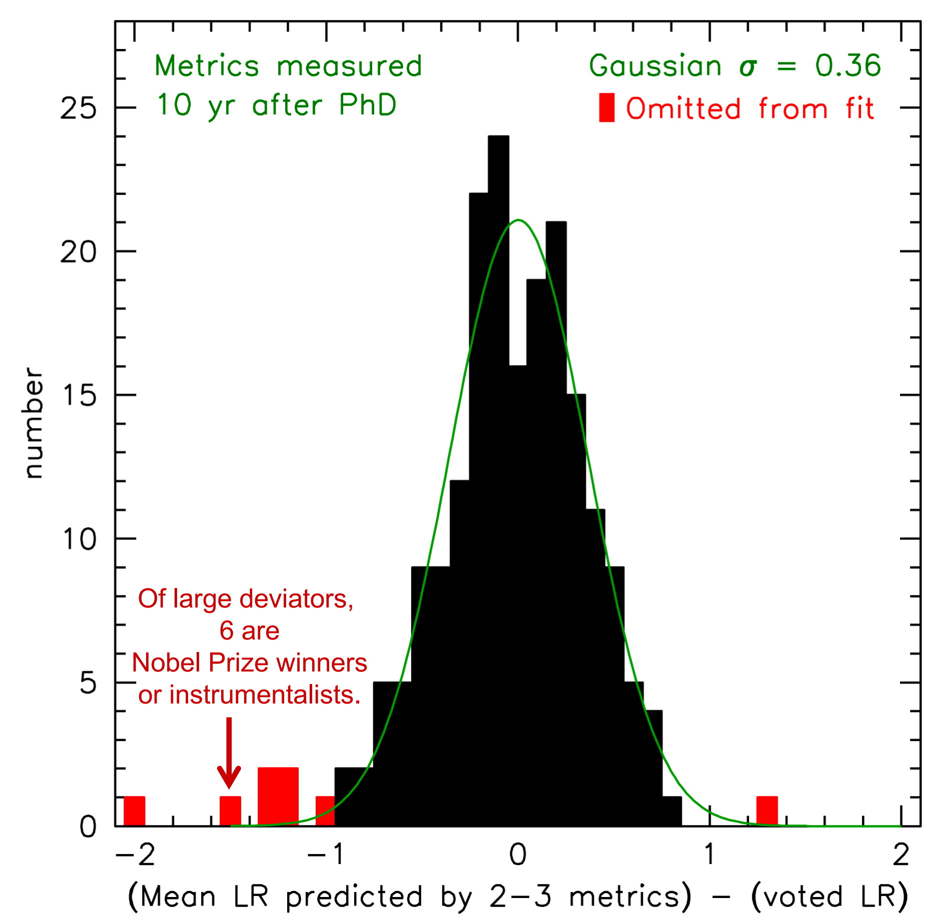


Typical single metric used as predictor has RMS = 0.37 (0.38).

Metrics of research impact in astronomy: Predicting later impact from metrics measured 10-15 years after the PhD

John Kormendy[a,b] (2021, Proc. Nat. Acad. Sci., in press)





Typical single metric used as predictor has RMS = 0.38.

Using <metrics> N yr post-PhD to predict future impact is a statistical tool with substantial uncertainties:

RMS(LR) is ~ 1/8 of the dynamic range.

But <u>all</u> opinions about candidates are statistical tools with substantial uncertainties.
Metrics add useful information to our judgments.