TTAR Memo No. 4

Gender-Related Systematics in the NRAO Proposal Review Process

Update Including all Proposals from Cycles 12A-21A

Gareth Hunt, Frederic R. Schwab and Patricia A. Henning
NRAO

10 February 2021

Abstract

This memorandum continues the analysis of gender-related systematics of the review of proposals submitted to the GBT, VLA, and VLBA originally published in 2016. After the publication of the original memo, the results were made known to the proposal reviewers. This was at a time when other astronomical observatories were analyzing gender bias in proposals to use their instruments. As a result of this, there is a strong movement towards dual-anonymous proposal evaluation system worldwide. Unfortunately, this cannot be practically deployed in our software and must await a new application. It is clear from the results presented here, that the gender imbalance is currently being ameliorated, mitigating concerns that any software deployment delay will adversely affect the issue of gender imbalance in the review process.

1 The NRAO observing proposal review process

Observing proposals are presently submitted at two deadlines annually, for observations to be made from February through July (cycle A) with a submission date of the previous August, and for August through the following January (cycle B) with submission in February. These are peer-reviewed for scientific merit by nine individual Science Review Panels (SRP) – eight SRPs until 2019B – each with at least six members. The scores are consolidated and are then assigned telescope scheduling priority considering primarily the review scores, but allowing for non-scientific factors such as LST pressure, by the Telescope Allocation Committee (TAC), which comprises the chairs of all SRPs.

2 Introduction

In 2016, Lonsdale et al. [LSH] undertook a study of gender systematics in the NRAO and ALMA proposal review processes. Reid [RHST] had previously performed a study on the same subject for the Hubble Space Telescope (HST) proposal system. Subsequent studies for the HST [JKHST] and ALMA [CALMA] have also been made. A significant gender-related effect (in favor of proposals with male PIs over those with female PIs) was found in the ALMA process, and a similar effect was found for the NRAO instruments, but to a lesser extent and with some reversals in the trend of male advantage, when examined by telescope and over time.

We have continued to monitor the gender systematics since the publication of that study. This note is a simplified update to include subsequent proposal cycles of the AUI/NRAO telescopes - the Green Bank Telescope (GBT), the Very Large Array (VLA) and the Very Long Baseline Array (VLBA). We do not here address subsequent ALMA reviews. We also do not address other potentially significant parameters such as PI seniority/prestige, geographic origin, and review panel science field.

The original paper [LSH] included results for the NRAO review process from 2012 (12A) through 2016 (16A). The first update [HSB] extended the review samples through 2019A (19A); this update extends the results through 2021A (21A).

3 Simplified presentation, Anderson-Darling

Since the scores are unmodified after the SRP peer-review process, we have used only the merged SRP (i.e., Normalized Linear Rank) scores for this analysis. The two distributions of scores awarded to female-PI and male-PI proposals were compared using several statistical tests (see Appendix A). In this section, we present a summary of the Anderson-Darling p-values (ADp). A low ADp indicates that the two distributions are not from a common parent distribution; a higher p-value is indicative of a degree of commonality. In order to compare the female-PI and male-PI distributions in the tables, we assign an imbalance key (IK) indicating which gender, if any, is favored:

- an **upper case letter (F** or **M)** if $ADp \leq 0.1$ (i.e., **very suggestive**). When there is an imbalance, the ADp does not indicate which distribution has the higher scores. This is determined readily by reviewing the data plots (section 4) to give **F** for imbalance in favor of proposals with female-PIs, **M** for male-PIs.
- a **lower case letter (f** or **m)** if $0.1 < ADp \leq 0.2$ (**suggestive**). The letter is determined as above.
- blank otherwise

Year	GBT	VLA	VLBA	All	%F SRP	%F TAC
2012					25.0	12.5
2013	F				26.0	18.8
2014	M			M	27.1	25.0
2015					21.9	6.3
2016		m		m	20.8	11.8
2017				M	26.5	18.8
2018	m	f			43.5	31.3
2019					45.9	64.7
2020					41.6	61.1
2021	f			F	43.6	55.6
All		m		M	32.1	30.6

Table 1: This is a simplified summary by year showing just the extreme cases of gender imbalance using an imbalance key (see section 3 for details). Both semesters in a year have been combined for the individual telescopes as was done in [LSH] for comparison with other telescopes, such as ALMA and the HST.

Our assignment of "suggestive" levels based on the ADp values is not rigorously justified, but is consistent with [LSH]; Babu and Feigelson [BF] recommend confirmation by bootstrap resampling, which has not been done for the Anderson-Darling analysis. However, the confidence intervals for our quartile score analysis of section 4 are computed directly via bootstrapping.

Tables 1 and 2 present a simplified summary equivalent to table 3 in the original paper [LSH].

It should be noted that most individual ADp values do not show a suggestive imbalance by semester or when combined to be presented by year. Note, for example in 2017, there was no imbalance for any individual telescope, but when considered together there was an indication of male advantage. Note also that the VLBA does not show any gender imbalance for individual semesters; the results are not statistically suggestive as the data are too sparse.

However, as mentioned in [LSH], the combination of all three telescopes can show a definite gender-related advantage; in years prior to 2018, there was an imbalance in favor of proposals with male PIs.

3.1 Reviewer gender ratio

[LSH] was published late in 2016, and the results were first presented to the reviewers for the 2017B semester. Guidance to the reviewers reminded them that the emphasis of the scientific reviews should be based solely on the scientific merits of each proposal. Since that time, overall imbalance, when it has been seen, has been in favor of proposals with female PIs. Since the publication of [LSH], the NRAO has implemented a policy requiring consistent effort on behalf of the recruiters to populate the SRPs with a gender ratio roughly equal to the astronomical community that we serve. This changing ratio is recorded in tables 1 and 2 and figure 1. Furthermore, the recruiters attempt to have at least two female reviewers on each SRP. Also included both in these tables and the figure is the gender ratio of the TAC. Although the TAC does not make changes to the SRP scores, it is the second level of the review process.

Semester	IK	ADp	%F PI	%F SRP	%F TAC
12A		0.86	26.9	25.0	12.5
12B		0.40	26.2	25.0	12.5
13A		0.34	25.5	25.0	12.5
13B		0.56	24.5	27.1	25.0
14A		0.51	29.7	29.2	25.0
14B	M	0.05	30.3	25.0	25.0
15A		0.73	24.8	18.8	12.5
15B		0.54	27.0	25.0	0.0
16A		0.65	30.2	25.0	12.5
16B	M	0.02	22.6	17.0	11.1
17A	m	0.12	32.7	20.8	12.5
17B		0.25	30.6	31.5	25.0
18A		0.22	32.1	38.9	25.0
18B	F	0.06	33.6	48.1	37.5
19A		0.73	32.0	48.1	50.0
19B		0.71	28.7	43.6	77.8
20A	f	0.13	36.5	39.3	66.7
20B		0.83	33.6	43.9	55.6
21A	F	0.10	31.9	43.6	55.6

Table 2: A summary by semester of the Anderson-Darling p-values comparing the distributions of the scores of proposals with female PIs and those with male PIs for all telescopes combined. Also included are the imbalance key (see section 3), the percentage of proposals with female PIs, and the percentage of female reviewers. The break is to indicate when changes in policies were adopted immediately following the publication of [LSH] (see section 3).

4 Graphical presentation, quartile score analysis

We modeled distributions of the quartiles of the normalized scores by bootstrap replication and resampling (see [E] section 5.3). A series of graphs and figures were generated for each year combining proposal cycles A and B for the individual telescopes and for the combination of all three. These graphs and figures are also generated for each semester for the combination of all three. The full set of these plots is available on-line.[1] In this section, we have extracted just the modeled distributions and combined them into four figures.

Figure 2 shows the quartile plots for each year, an extension of figure 2 in [LSH], for 2012 to 2016 and this continues here in figure 3 for 2017 to 2021A. Note that all early years show a higher score distribution for proposals with a male PI; 2021A is the sole exception, consistent with the IKs in table 1.

Figure 4 shows the same information for semesters 2012A-2015B, figure 5 for 2016A-2019A and figure 6 for 2020A-2121A. Although many semesters appear to have a higher score distribution for one gender or the other, only six are statistically suggestive as seen in table 2. Finally, figure 7 shows the aggregated periods 2012A-2017A and 2017B-2021A.

Since the Anderson-Darling method matches the complete distribution, giving higher weight to the edges of the distribution, the results compared to other methods are not necessarily intuitive.

5 Conclusions

The results reported here, extending the analysis of [LSH] and [HSB] to semester 2021A, show that the outcomes of the NRAO proposal review process tended to favor proposals from male PIs over those from female PIs before 2018, but not since then. There are insufficient data to identify trends rigorously, but

[1] https://science.nrao.edu/science/reports/StatisticalData

the indications look promising that the problems have been addressed (see figure 7), with ADp/IK values of 0.01/"M" and 0.11/"f" before and after the publication of [LSH].

AUI/NRAO is concerned by the gender-related imbalance revealed by these studies. AUI/NRAO is committed to a fair and equitable proposal review and time allocation process, and actively emphasizes to all reviewers each cycle that rankings and decisions must reflect only scientific merit, technical feasibility, and operational constraints. Since 2017, the NRAO has worked to raise awareness of these issues with reviewers and its user community, and has successfully achieved more balanced gender representation on its review panels. The Observatory is committed to delivering greater fairness in future, and is monitoring developments being implemented elsewhere.

Dual-anonymous reviews (anonymous authors and PIs plus anonymous reviewers) are becoming the standard in astronomy. This is a reasonable path to follow, and NRAO intends to implement a dual-anonymous review process in future. A replacement software suite is currently under active development, but it will not be available soon. It is clear, however, from the results presented here, that the gender imbalance is currently being ameliorated, mitigating concerns that any software deployment delay will adversely affect the issue of gender imbalance in the review process.

6 Disclaimer

This note includes much unmodified text from its predecessor [HSB] so that all relevant information may be found in one location. Data may have changed slightly because a few additional gender identifications are now available.

7 References

LSH Carol J. Lonsdale, Frederic R. Schwab and Gareth Hunt, *Gender-Related Systematics in the NRAO and ALMA Proposal Review Processes*, http://arxiv.org/abs/1611.04795, 16 November 2016.

HSB Gareth Hunt, Frederick R. Schwab and Lewis Ball, *Gender-Related Systematics in the NRAO Proposal Review Process Update Including all Proposals from Cycles 12A-19A*, NRAO TTA Memo #3, 25 February 2019.

RHST I. N. Reid, *Gender-based Systematics in HST Proposal Selection*, PASP, 126, 923, 2014.

JKHST Stephanie K. Johnson and Jessica F Kirk, *Dual-anonymization Yields Promising Results for Reducing Gender Bias: A Naturalistic Field Experiment of Applications for Hubble Space Telescope Time*, https://ui.adsabs.harvard.edu/abs/2020PASP..132c4503J/abstract, 2020

CALMA John Carpenter, *Systematics in the ALMA Proposal Review Rankings*, https://ui.adsabs.harvard.edu/abs/2020PASP..132b4503C/abstract, 2020

BF G.J. Babu and E.D. Feigelson, *Goodness-of-fit and all that!*, Astronomical Data Analysis Software and Systems XV, ASP Conference Series #351, 127, 2006.

E Bradley Efron, *The Jackknife, the Bootstrap and Other Resampling Plans*, Society for Industrial and Applied Mathematics, 1982.

A Statistical tests

There are several widely accepted tests to compare statistical distributions. Among these are Cramér–von Mises, Kolmogorov–Smirnov and Anderson–Darling. These have been traditionally used to test the null hypothesis that distributions are from different populations when the p-value is small. Nevertheless, it is noticeable that higher p-values do provide a suggestive indication of degree of difference.

We compute the statistics comparing the observed gender distributions for each of the cycles or combination of cycles mentioned in the main text. The imbalance keys (introduced in Section 3) are derived just from the Anderson-Darling p-values. We have made comparisons with other distribution tests, and we are confident that using just the Anderson-Darling test is adequate.

For example, we analyze the complete ensemble of data for Cycles 12A through 17A (GBT, VLA, and VLBA), and compare this with other combinations of cycles. Among all of the AD p-values in our summary, the value 0.0084 for Cycles 12A-17A is the very lowest. Similarly, the 12A-17A Kolomogorov-Smirnov p-value, 0.0357, is lowest; and the Cramér-von Mises value, 0.0257, is second-lowest (the lowest is 0.0191 for 16B).

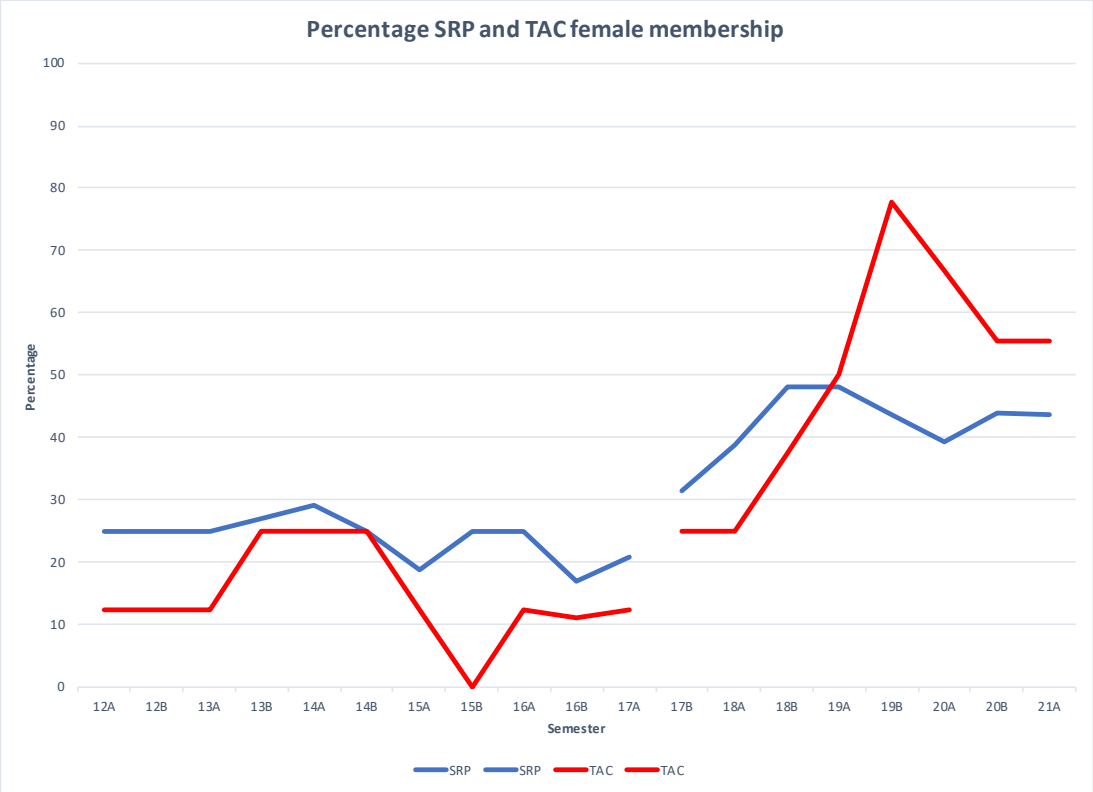

Figure 1: The female fraction of the SRP and TAC by semester. The break is to demonstrate the change in the SRP gender ratio after the publication of [LSH] and the policies adopted immediately thereafter (see section 3).

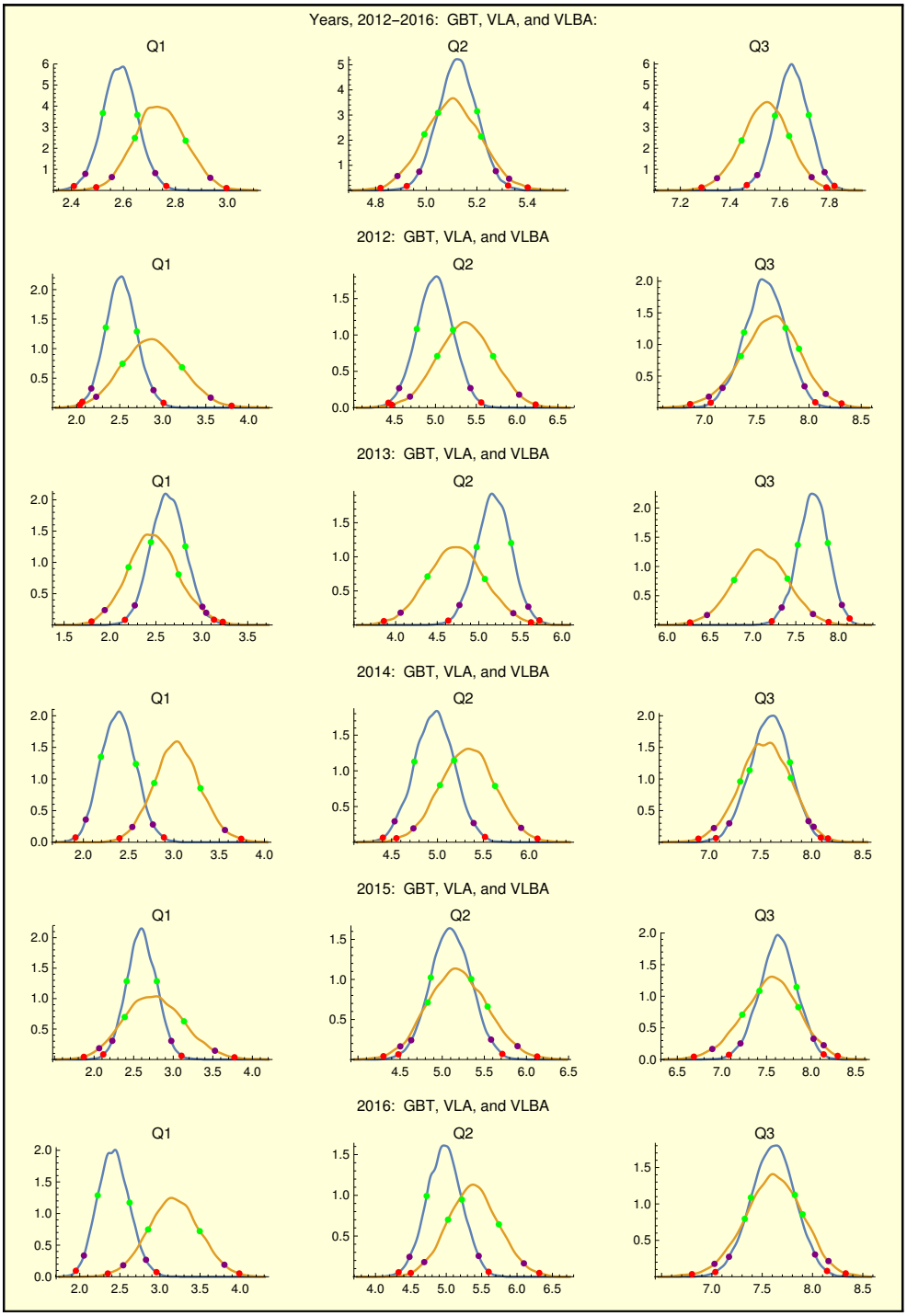

Figure 2: Modeled distributions of the 25th, 50th and 75th percentile of the normalized scores of proposals submitted to all telescopes derived by bootstrap resampling. Orange curves: proposals with female PIs; blue curves: male PIs. Green, purple and red dots delimit, respectively, the 68%, 95% and 99% probability intervals. Top: total; subsequent rows are for years 2012-2016.

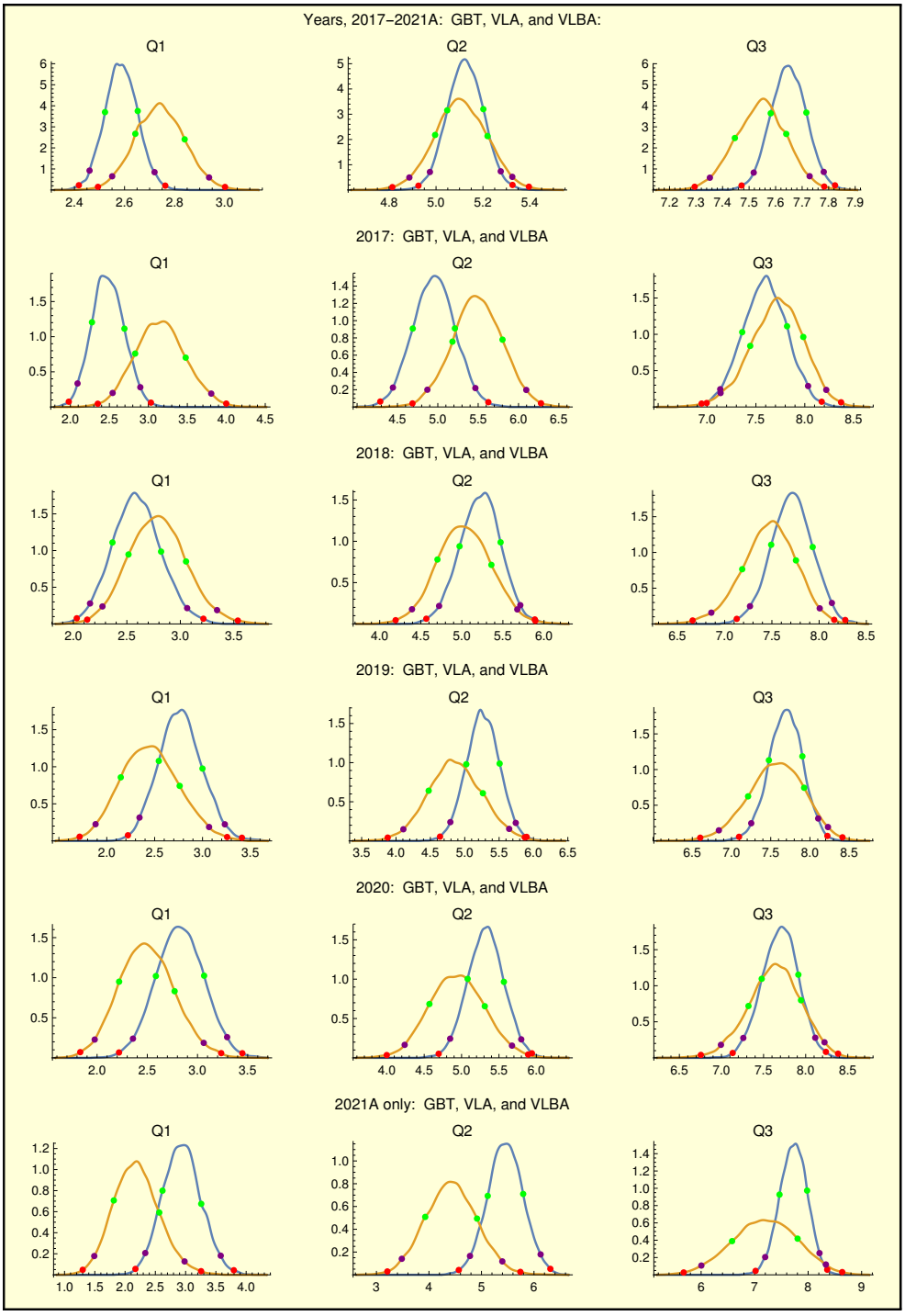

Figure 3: As for Figure 2. Years 2017-2021A.

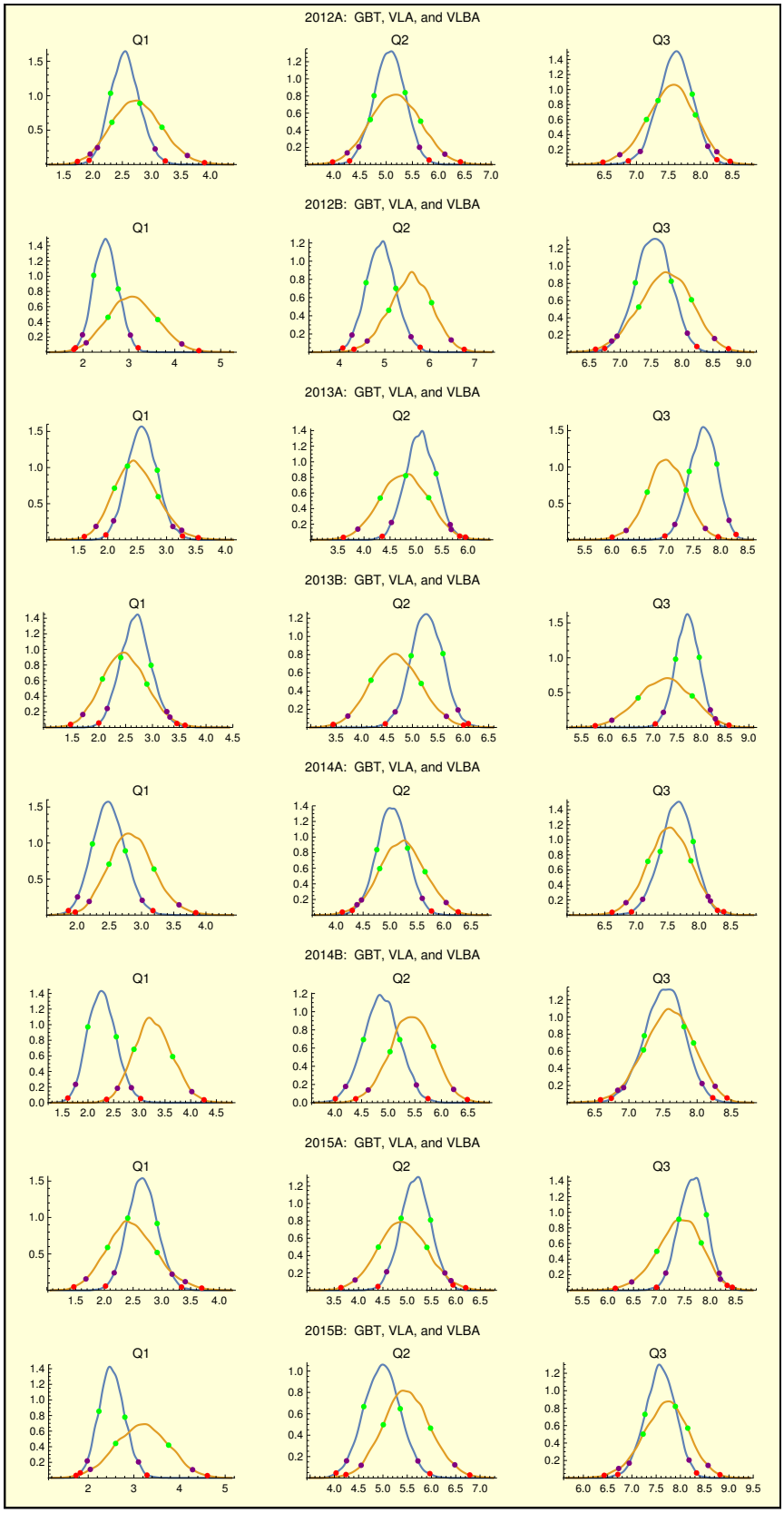

Figure 4: As for Figure 2. Semesters 2012A-2015B.

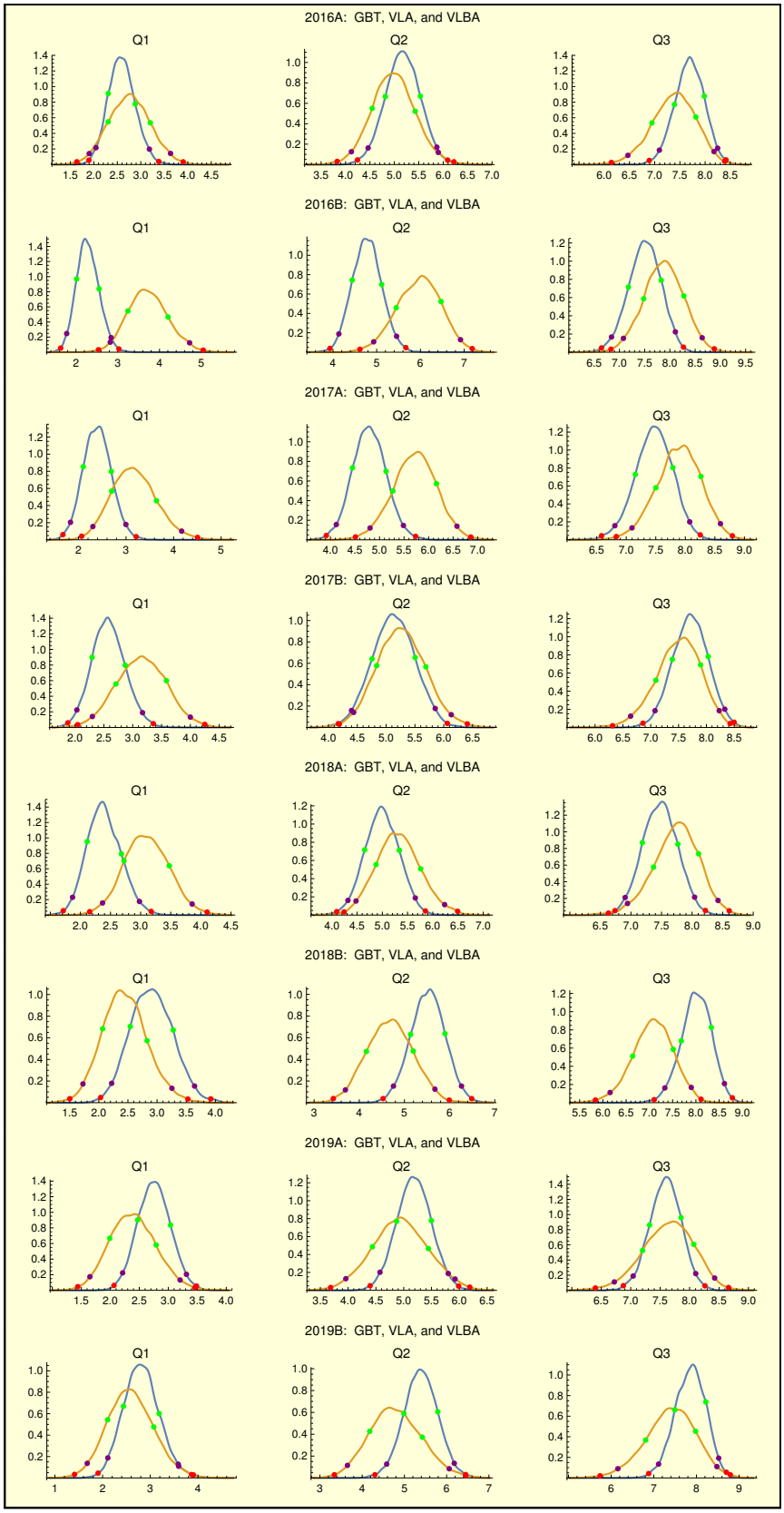

Figure 5: As for Figure 2. Semesters 2016A-2019B.

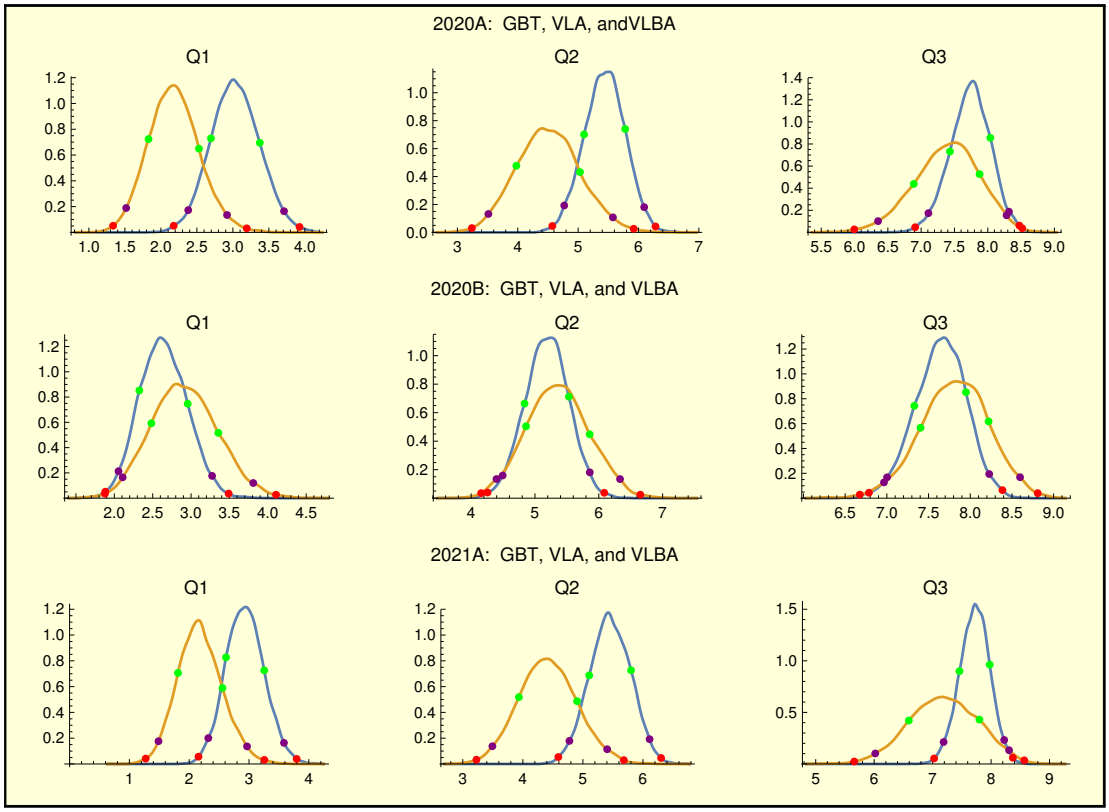

Figure 6: As for Figure 2. Semesters 2020A-2021A.

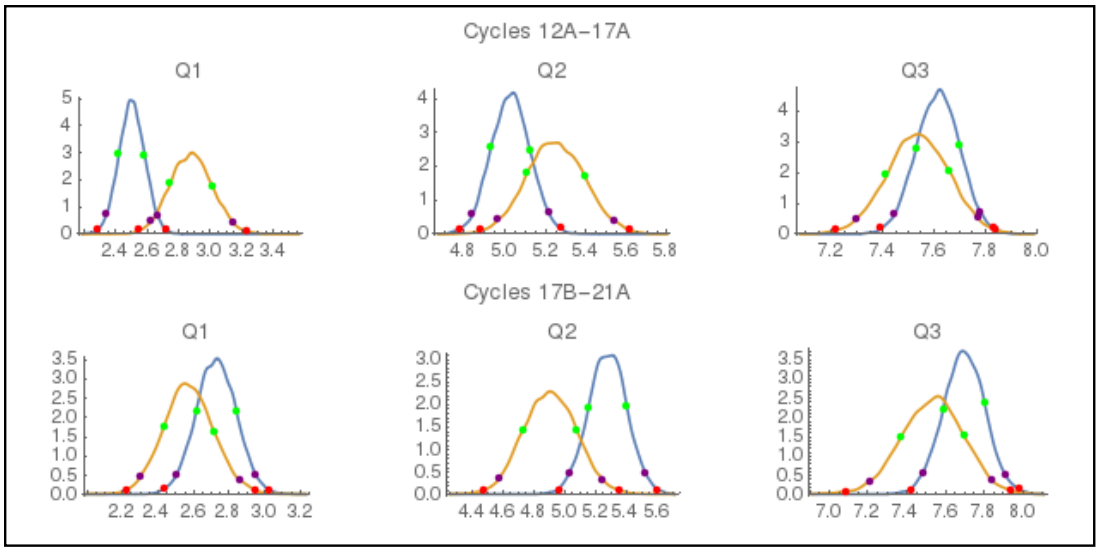

Figure 7: As for Figure 2. Distributions for the two periods 2012A-2017A and 2017B-2021A. These show the very suggestive improvement before and after the implemented changes as a result of [LSH]. The corresponding AD p-values are 0.0084 "M" and 0.1141 "f" respectively.