To the referee and editor,

The authors wish to thank the referee for his or her valuable comments in the improvement of this paper. As a direct result of the referee's involvement, we have constructed a more complete analysis of the topic at hand and are confident in the contributions our work has to offer the current state of the field. The trends we qualitatively noted in our original manuscript now have much more substantial evidence as a result of the work that has been conducted in response to the referee's comments, and we wish to convey our gratitude for their deliberate and thorough response to our research.

Referee Report
Reviewer's Comments:

In this manuscript, the authors compare samples of galaxy-galaxy lens candidates identified in the same fields but by three different methods (spectroscopy, machine learning, and citizen science). Their main finding is that there is remarkably little overlap between the three samples, which the authors attribute to the parent samples spanning different redshift ranges or the methods being sensitive to different imaging configurations. However, considering that the three sets of lens candidates are selected from three distinctive parent samples, I think such little overlap is well expected. In my opinion, the findings presented in the current manuscript are some simple observations that are well recognized in the community, and are not meaningful enough by themselves to be published as a separate paper. Also, I can see that their conclusions may have been biased because lens candidates are getting compared. In addition, I notice some obvious mistakes. I think a major revision is needed before this manuscript can be re-considered, and I encourage the authors to carefully proofread the manuscript.

Major comments

>>> By design, the spectroscopy method requires at least some portion of the lensing features being close to (and likely blended with) the lensing galaxies, while image-based methods such as machine learning preferentially select lens candidates with lensing features reasonably separated from the lensing galaxies. So the little overlap between these two methods is trivial and straightforward to understand (e.g. see Sonnenfeld et al. 2018).

<<< ***The focus of this paper is to understand the biases that result in the lack of overlap between methods applied to the same overall parent sample, with the intent to maximize the potential of these methods to retrieve a more complete sample of identifiable lenses offered by a particular region of sky. We have reassessed the data as a whole and followed several specific suggestions offered by the referee (to be detailed later) and report our conclusions in a more detailed and evidence-supported way. We thank the referee for their contributions to the evolution of this paper from its original form.***

>>> My another biggest concern is, the current comparisons do not directly reflect the intrinsic selection biases of the three methods themselves. Also, conclusions drawn from comparing lens candidates, especially ones that have not been reasonably cleaned, can be further biased and even wrong!

*<<< **Additional cleaning of samples and analysis of parent samples and selection biases have been included and are detailed in more specific terms in response to later referee comments. Several sections have been added and drastically revised in order to achieve this. We critically assessed our work in response to the referee's report and have made significant changes to the manuscript.***

>>> 1. To quantify the selection biases of the three methods, I think it's necessary to compare lens candidates with the parent sample from which they were selected for each individual method independently instead of directly comparing the three end products! For example, Bolton et al. 2008 compared SLACS A-grade lenses and lens candidates with a control sample, and showed using KS tests that the SLACS lenses and lens candidates are statistically consistent with the parent SDSS galaxy population in terms of stellar mass, size, and velocity dispersion. Sonnenfeld et al. 2018 used two image-based methods and one spectroscopy method to select lens candidates and compared their performances. Their comparison is similar to what's done here, but more sensible because their three methods used the same parent sample. They found that lenses and lens candidates are slightly biased towards the high-mass end compared to the parent sample, but no particular redshift region is favored.

<<< The ultimate parent sample for each method is the same: the GAMA equatorial regions and KiDS imaging. These surveys are closely related and cooperate in a number of ways, including the GAMA-KiDS GalaxyZoo effort discussed in the paper. These equatorial regions represent a parent sample identical to both, one focused on spectroscopic completeness while the other is focused on high-fidelity imaging to the same depth. We appreciate the author's note and hope we have clarified this in the manuscript.

We also conducted K-S tests between each candidate catalog and the direct "parent subsample" on which the selection technique was applied. These tests show a significant difference between naive parent samples and selected lenses, and a new section (Section 5. Selection Effects) has been added to discuss this. This was a valuable course of inquiry that we thank the referee for suggesting.

>>> 2. The little overlap between Petrillo et al. (2018a) and GalaxyZoo does not necessarily mean "these low masses are ... well below the range of identification by the other two methods". I suspect that many GalaxyZoo candidates are actually low-mass, blue galaxies that were simply skipped (and thus unclassified) in Petrillo et al. (2018a) because Petrillo et al. (2018a)

only looked at luminous red galaxies (LRGs) in the first place. In principle, the machine learning method itself should be able to largely recover candidates from citizen science as they're both based on imaging data. To make a fair comparison, I suggest either singling out GalaxyZoo candidates that also pass the LRG cuts used in Petrillo et al. (2018a) or applying the machinery in Petrillo et al. (2018a) to all GalaxyZoo candidates.

*<<< **In principle, we agree, and we have added additional statements to clarify our assertion that in practice machine learning can and must take steps to improve upon this bias. The initial choices in both GalaxyZoo and LinKS ML made a mutual agreement unlikely. This was our point: the sample selections designed out the possibility that these two methods could be used to verify each other or to construct a complete census of identifiable strong lenses within the surveyed area. Following color-magnitude selection of LRGs as outlined in Petrillo et al. 2018, only 6 of the 36 GZ candidates and 7 of 47 ML candidates passed the selection criteria. Description of the color-magnitude selection is given in the new Section 5, and Subsection 6.4 of the Discussion presents the results of application of that selection post-identification to the candidates.***

>>> 3. The fact that there are more overlaps between machine learning and citizen science when the lens score threshold is lowered seems to suggest that many GalaxyZoo candidates are simply not good lens candidates, perhaps because they were selected with less physical constraints compared to the machine learning method. For example, galaxies with obvious, widely-separated arcs would have a high "Lens or Arc" score in GalaxyZoo, but they are simply not lenses given the separations and masses. Again, comparing lens candidates, especially ones that have not been reasonably cleaned, can be biased! It would be useful to provide the separations between the arcs or arcs and galaxies for the GalaxyZoo candidates and their estimated Einstein radii based on velocity dispersion (from spectroscopy) and some fiducial source redshifts.

*<<< **Upon the referee's suggestion, we reassessed sample selection of all three methods, introducing additional objective criteria to remove obvious false-positives. The revised manuscript includes additions to Section 3, Data and Observations, and an additional subsection detailing the estimation of Einstein radii for all candidates. Those estimates are shown in several plots, and we explore the relationships between them and redshift/stellar mass of the candidates, as well as comparing them to other surveys (SLACS, S4TM, and BELLS).***

Increased purity is achieved by taking:

1. ***GAMA spectroscopy candidates with redshift differences $\Delta z > 0.1$ between lens and source.***
2. ***LinKS machine learning candidates with scores above the threshold of 17, from Petrillo (2017).***

3. ***GalaxyZoo candidates with the same scoring criteria but an additional separation of Einstein radius greater than the PSF of the KiDS survey.***
    a. ***We also note that high-scoring candidates that do not pass this latter selection criterion may be useful for other galaxy studies (eg. tidal features).***

We maintained the results from our initial submission (moved to the Appendix) to show that relaxed vetting introduces more noise along with its additional overlaps, affirming that the distinctness of the candidate samples is not due to overselection on our part. Naively, one would assume easily identifiable lenses in any survey should be easily identified with any method, but that is not the case. We thank the referee for challenging us to reassess the data in a more rigorous fashion.

4. The spectroscopy sample needs to be cleaned as well. I checked the PG+ELG sample in Holwerda et al. 2015, and found that the redshift differences between the foreground and background galaxies are small for the majority of the PG+ELG pairs. Almost 50% of the PG+ELG sample have redshift differences smaller than 0.1, and almost 70% have redshift differences smaller than 0.2. I actually highly doubt that those PG+ELG pairs with small redshift differences could be true lenses given the seemingly small lensing cross sections. Of course, it would be very interesting if some of the low-mass candidates in Holwerda et al. 2015 turn out to be true lenses. But I suspect that the false positive rate in this spectroscopy sample is much higher towards the low-mass end. Then it's not surprising to see the spectroscopy candidates appear less massive on average than the machine learning candidates. It's again worth estimating the Einstein radii and lensing cross sections for the spectroscopy candidates from velocity dispersions and redshifts. The authors can also estimate the lensing probabilities for the spectroscopy sample following the procedures in Bolton et al. 2004, which should give a better idea of how many true lenses are expected from this sample.

<<< ***We instituted the additional selection criteria to Section 3.1 as described in the previous comment. Several of the lower-mass candidates remain and are worthy of follow-up. This was a useful exercise, and we thank the referee for their suggestion.***

Intermediate comments:

Section 2
>>>Nearly 100 strong lenses have been confirmed in the SLACS survey (see Auger et al. 2009).

<<< ***As of Auger 2009, SLACS has discovered 85 confirmed lenses. This has now been addressed.***

>>> For the SLACS survey, approximately 50%, instead of "nearly 100%", of the candidates with HST observations were confirmed as lenses.

*<<< **This has now been addressed.***

>>> I don't think "the lensing arc does not intersect with the SDSS aperture" was the main reason that most SLACS strong lenses are at z > 0.1. First of all, the lensing cross section drops rapidly below z~0.5 (e.g. Hilbert et al. 2008, Lapi et al. 2010), so naturally there are fewer lenses at lower redshifts. In addition, the SLACS survey (Bolton et al. 2006) imposed a redshift cut of z > 0.15 for the majority of the lens candidates that were selected from the LRG sample.

*<<< **We have added a reference to Sonnenfeld-2015 (and references therein), which claims that "the Einstein radius seems to be the main quantity determining the detection probability. Of the two terms in the selection function, the Einstein radius selection is the dominant one while the lensing cross section correction has little effect on the results of our analysis." They describe a limit on Einstein radius: "lenses with too large Einstein radii can escape the selection because the lensed features contribute little to the flux deposited within the 1."5 radius fiber used by SDSS spectroscopic observations." Einstein radius is a function of the mass and redshift arrangement, with higher mass and closer lenses resulting in larger Einstein radii, so for a given positioning of lens and source there is a soft upper threshold for the amount of mass that will result in an Einstein radius whose light will contribute significant flux to be identified by a spectroscopic aperature of fixed angular size. We note both arguments and references in the paper and calculate these maximum masses, discussed in Section 6.2 of the Discussion and shown in Figure 10.***

Section 2.1
>>> The authors claim that "The spectroscopic approach typically results in a small but very clean sample". Given the typical 50% success rate of the SLACS survey, I'm not sure whether I'll call this "very" clean. One can probably get a similarly small and clean lens sample using the machine learning approach (by adopting a high score).

*<<< **This has now been addressed by new selection criteria detailed in a previous comment.***

>>> Again, I don't agree "spectroscopy structurally misses lower redshift lenses". I think this selection bias is very mild, and only at the lowest redshifts (z ~< 0.1).

*<<< **We address this in discussion of the limits based on spectroscopic fiber aperture size in the manner described above.***

Section 3.1

>>> As mentioned previously, there seems to be quite some false positives in the spectroscopy sample. I would be very cautious about calling "blended spectra provided arguably the cleanest dataset compared to the others" and "the 85 candidates are ... reliable" unless it's been tested.

*<<< **This has now been addressed. We appreciate the challenge to reassess our assumptions about this particular set of data.***

Section 3.3
>>> I don't think Figure 6 shows "The cuts we imposed do not skew the results ...". I would expect to see all GalaxyZoo candidates in Figure 6 and comparisons of the redshift and stellar mass histograms.

*<<< **GalaxyZoo selection has been readdressed and presented more clearly in comparison to other data.***

Section 4
>>> It's clearly not the methods that "identified galaxies with a distinct range of redshifts", but the parent samples are selected that way.

*<<< **K-S tests of candidate samples against parent samples (with description of parent sample construction) have been included and show differences between candidate samples and the parent subsamples from which they were selected. This is discussed in new Section 5 (Selection Effects).***

>>> I don't quite understand the statement of "there is a maximum total mass ...". I don't think there is a sharp mass limit for the spectroscopy method. Please clarify.

*<<< **As described above, we have included additional references, calculations, and plots to address this particular point. We hope we have addressed and successfully clarified these points.***

>>> The authors seem to indicate that the spectroscopy method tends to find lower-mass lens candidates compared to the machine learning method, which is not well supported in my opinion. In the literature, Sonnenfeld et al. 2013 actually showed that the stellar mass distribution of the SLACS sample is similar to the SL2S sample, which were selected based on imaging data and should be statistically comparable to a machine learning sample. As I mentioned previously, Bolton et al. 2008 showed that the SLACS sample is statistically consistent with its parent sample in terms of mass. And it's possible that the majority of the low-mass spectroscopy candidates in Holwerda et al. 2015 are not lenses.

*<<< **The result here is not intended to describe the results of mixed spectroscopy as a general method; rather, it shows a bias in both methods when applied to a survey with different characteristics. The difference between the GAMA candidates and SLACS (upon***

which the machine learning sample is trained) is precisely what we are showing. GAMA finds smaller candidates because of its increased depth and completeness and does NOT return the higher mass candidates of SLACS and Petrillo's machine learning because higher-mass lenses tend to have larger Einstein radii. The result is that there is insufficient source flux detected by the smaller GAMA fiber compared to the SDSS fiber. Where appropriate, "spectroscopy" has been changed to "GAMA spectroscopy" to reflect the focus of the statement. We thank the referee for pointing out this ambiguity.

Section 4.1
>>> Why was G136604 not identified in Holwerda et al. 2015? It's worth showing the SDSS and GAMA spectra for this system. A related question is, how many SLACS+SLACS for the Masses+BELLS lenses are in the GAMA spectroscopic sample, and how many are recovered in Holwerda et al. 2015?

<<< ***The Einstein radius of G136604 is larger than the 1" radius of the GAMA aperture, so insufficient source flux is present for detection. SLACS, S4TM, and BELLS Grade-A lenses that correspond to candidates within the equatorial GAMA fields have been considered and are shown (including G136604); none were recovered by the GAMA spectroscopic method. We thank the referee for the useful suggestion.***

>>> It's worth showing the KiDS image alongside the HST image for G593852 in Figure 10.

<<< ***We believe the comparison between KiDS and HST imaging is clearly shown by the images already included in the paper. G593852 shows lower contrast than the other three images because the F625W filter is significantly narrower than F814W. This note has been added to the paper for clarity.***

Section 5.1
>>> It's actually interesting to show the KiDS images for the 5 overlapping spectroscopic lens candidates when more relaxed cuts are used, and highlight them in Figure 8. Even though they don't have a high lens score according to imaging data, the fact that they also show higher-redshift emission lines makes them highly probable lenses.

<<< ***Additional selection criteria have removed these from consideration.***

Section 5.2
>>>It's better to demonstrate using a figure that "the candidates it identified are characteristically similar ... to those identified by SLACS spectroscopy".

<<< ***We have shown in Figure 8 that the Einstein radius estimates are characterstically similar. We also note that the intent in the LinKS selection was to recover SLACS lens candidates, and their internal study shows the similarity with SLACS candidates.***

>>> The statement of "the resulting identified lenses are equally biased toward higher masses" needs to be tested.

*<<< **Plots and K-S tests reveal bias toward higher mass in the LinKS selection in comparison with the parent subsample (determined by LRG color-magnitude cuts) and in comparison with candidates identified by other methods.***

Section 5.4
>>> The SLACS and SLACS for the Masses lens samples were selected from the SDSS LRG sample that has a limiting magnitude of r=19.5 (Eisenstein et al. 2001) instead of what's shown in Figure 12. The BELLS lens sample was selected from the BOSS LRG sample that also has a limiting magnitude of r~19.5 (Eisenstein et al. 2011).

*<<< **According to Eisenstein 2001, the SDSS Main sample has a limiting magnitude of r < 17.7, which led to our misunderstanding. Eisenstein 2001 says "the sensitivity of SDSS in g is sufficient at r ~ 19.5 to yield a well-measured g−r color, and one can combine this with r−i to constrain the redshift independently of the SED of the galaxy." We take r < 19.5 to be the effective limiting magnitude for the LRG sample, and update our analysis of these surveys in Figure 15. We appreciate the referee's identification of our mistake.***

>>> The limiting r-band magnitude and mean seeing for the DES sample is 24.9 mag and ~1.1 arcsec according to Jacobs et al. 2019. The typical seeing for the DECaLS sample is approximately 1 arcsec according to Huang et al. 2019.

*<<< **These numbers have been corrected in the analysis of these surveys and presented again in Figure 15. We again appreciate the referee's diligence in checking our mistake.***

>>> It's unclear from the color bar in Figure 12 what the candidates densities are. Please also provide the actual calculations and values in the text. For SLACS, SLACS for the Masses, and BELLS, is the number of confirmed lenses or the number of lens candidates used? I'm surprised to see that GAMA II has significantly higher candidates density than the other three spectroscopy samples that have similar limiting magnitudes. What makes GAMA II so effective? Or maybe lots of GAMA II candidates are simply not lenses?

*<<< **For all surveys, the number of high-quality candidates is taken against the on-sky angular survey size to acquire candidate densities. We attribute GAMA's effectiveness to its completeness, ie. the number of spectra taken per unit area of sky, as shown in Figure 15.***

>>> Please add "candidates" after "Figure 12 shows numbers of identified lenses". It's really important to emphasize that these numbers are only candidate densities, especially for GAMA II, KiDS, DES, and DECaLS.

*<<< **We recognize this point as being essential to the clarity of the entire paper and appreciate the referee's diligence. We have changed the language throughout the paper to clarify the distinction of lens CANDIDATES versus confirmed lenses.***

Section 6
>>> The first two conclusions are not precise. Please see the above comments.

*<<< **The conclusions have been reassessed and are supported by the evidence added in response to the referee's comments. We thank the involvement of the referee for the challenge to refocus and arrive at more precise conclusions.***

>>> The statement of "Blended spectra identifies only ..." is incorrect. As long as some lensing features fall in the aperture and are bright enough, they would be identified.

*<<< **This statement has been amended, clarified, and supported by reference material, calculations, and plots presented in the manuscript.***

>>> I actually think that the fundamental limitation of the machine learning approach in Petrillo et al. (2018a) is imaging resolution rather than the training set. For a particular resolution, even if the training set extends to lower masses, Petrillo et al. (2018a) will still preferentially select candidates with lensing features that can be clearly separated from the lensing galaxies, which naturally results in a sample that is biased towards more massive galaxies.

*<<< **We note the preferential selection of all three methods by analyzing properties of the systems, estimating Einstein radii, and performing K-S tests. We include references to Li-2020, which recovered twice as many candidates using the LinKS methods by experimenting with ignoring the assumption of LRG colors. We believe that further exploration of machine learning algorithms away from this assumption and the size assumption will better prepare these techniques for application to future surveys that will improve the resolution limitation to a large degree.***

Minor comments:

Section 2
>>> Please follow the naming conventions in the discovery papers. SLACS stands for "Sloan Lens ACS", and SLACS4MASSES should be either SLACS for the Masses or S4TM.

*<<< **This has been addressed throughout the paper. We appreciate the correction.***

Section 2.1

>>> The authors mention that "Holwerda et al. (2015) identified 102 strong lensing candidates", but there are 104 lens pair candidates in Holwerda et al. (2015). Please explain why two are missing.

<<< ***This typo has been rectified. We thank the referee for close inspection.***

The authors seem to imply that ~70% of the spectroscopic lens candidates are true lenses by using "verified", while in fact Chan et al. 2016 only found 10 probable lens systems among those candidates. Please rewrite this sentence.

<<< Chan

Section 3
>>> Are redshifts for the machine learning catalog and citizen science catalog also spectroscopic? What is the typical uncertainty in stellar mass? It would useful to include error bars in the plots involving stellar mass.

>>> ***All redshifts are spectroscopic. Mean uncertainty in stellar mass of the presented data has been added to Figure 9. We thank the referee for the useful suggestion.***

Section 3.3
>>> The solid line in the Lens vs. None panel of Figure 5 does not look like a 2:1 line.

<<< ***This figure has been revised.***

>>> References with arXiv numbers need to be updated if possible.

<<< ***This has been addressed.***

–

The authors wish to repeat our gratitude for the efforts and diligence of the referee, whom we believe has had an extremely valuable effect on the direction, rigor, and focus of this research. We appreciate the referee's commitment to offering their valuable time to the improvement of the work and hope we have sufficiently addressed all of their points in this revised manuscript. Thank you for your work.

- Shawn Knabel and collaborators