We thank both referees for the careful study of our paper and the well thought-out, helpful comments to improve it. Our answers follow below.
(Referee comments are in blue, our answers in black, text modifications are slanted)

Answer to Referee A:

The manuscript reports a first measurement of averaged elastic electron-proton and positron-proton scattering differential cross sections, which are expected to be insensitive to the leading effects of hard two-photon exchange (TPE). The TPE has been a topic of great interest in the electron scattering community in the last two decades motivated by the large discrepancy observed in the proton electromagnetic form factor ratio extracted between experiments measuring the recoil proton polarization versus those measuring the unpolarized scattering cross sections. A number of experiments specifically designed to test the TPE predictions and to resolve this discrepancy have been conducted in recent years, however, the results are still far from being conclusive. The reported work is taking a different angle, i.e. by minimizing the hard TPE effect in lepton scattering in order to access the proton electromagnetic form factors in a more robust way. Therefore, the reported results are quite interesting, and timely.

However, before the paper be considered further whether it is suitable for a PRL publication or not, a number of areas listed below need to be improved:

1. The introduction of the paper as currently written has not taken into account some of the latest developments in the field concerning the relationship between the nucleon's electromagnetic form factors and the charge and magnetization distributions inside. The introduction is also somewhat focused more towards a nuclear/hadronic physics audience.

We reformulated our introduction, it now reads:
"As the lightest stable composite particle emerging from quantum chromodynamics, the proton is one of the best testing grounds for our understanding of the strong force. One of the ways of characterizing the proton's internal quark-gluon structure is through measurements of elastic electron-proton scattering, from which the proton's electromagnetic form factors, GE and GM, can be extracted. These form factors reveal information about how electric charge and current are distributed within (though this relationship is far from simple, see [1] (Miller, 2018), and provide a touchstone for the verification of theoretical descriptions and computational approaches."

2. Page 2, last paragraph in the left column on the Monte Carlo simulation, I quote: `` Acceptances, radiative corrections and efficiencies were accounted for via a sophisticated Monte Carlo (MC) simulation, which matched the measured time-dependence of the beam current and position, rigorously treating the correlations between effects." This passage does not make sense. Appears to be an unfinished cut and paste job.

We reformulated this sentence to now read:
"Acceptances, radiative corrections and efficiencies were accounted for via a realistic Monte Carlo (MC) simulation. The MC parameters, for example beam position and beam current, were adjusted dynamically to match the values recorded by the slow control system to simulate the time-dependence of these quantities. This approach also rigorously captured the possible correlation between parameters."

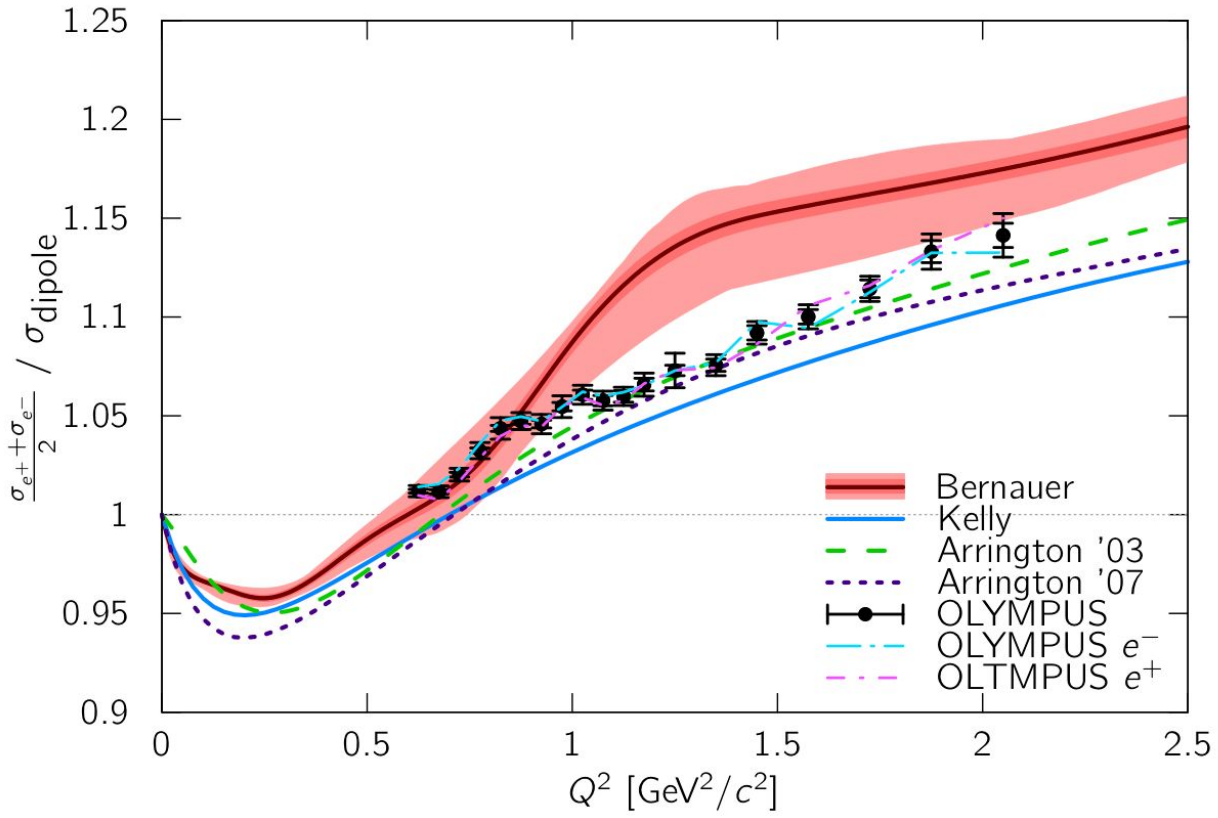
3. Page 2, near the end of the right column, ``Four independent elastic event selection routines" were mentioned without any description at all. Although four references were given, they are three PHD theses plus an unpublished technical note. I do not think this is sufficient and the authors need to provide a concise description of each of the four event selection routines.

With the limited space, we can unfortunately not describe the analysis in much detail. We expanded the section slightly to express a little bit information about the differences:

"Four independent elastic event selection routines were developed [35–38], which allowed us to assess the degree of event-selection bias. While the four approaches differ in detail, they all exploit the fact that, for a coincidence measurement of elastic scattering, the kinematics are over-determined and that selection cuts on the self-consistency of the kinematics can be used to suppress inelastic background. The routine of Ref. [37] used wide cuts in lepton-proton vertex time correlation, vertex-position correlation, polar-angle correlation, and momentum correlation to reduce inelastic background, before estimating any remaining background using side-bands in the azimuthal distribution of track pairs. The routine of Ref. [36] performed different selection cuts and is distinguished by performing particle-ID at the level of track pairs, rather than individual tracks. The routine of Ref. [35] examined background over a two-dimensional space of polar and azimuthal angle correlation. The routine of Ref. [38] built an elastic-pair probability for track pairs based on their vertex, time and angle correlations, and the missing energy assuming elastic kinematics. Low probability combinations were rejected. The surviving best pair for each event were then used for the rate extraction, with a background estimate based on the coplanarity in Q^2 slices."

4. In Fig.2, it will be interesting to see the individual results from e-p and e+-p together with the charge averaged results.

We would have liked to do this, but the problem is that the points are so close that they'll overlap the error bars. The best solution we found is attached here, but we still feel that it's not easy enough to read.



5. I would argue that what's shown in Fig.3 supports my request in #4 above, which is to show the e-p and e+p results separately, in addition to the average cross sections.

Prior to radiative corrections, the electron and positron yields are different (by as much as ≈7%, as Fig. 3 confirms) but this reflects a combination of known differences in radiative effects (soft TPE, bremsstrahlung interference) convolved with the OLYMPUS spectrometer acceptance. After correcting for these effects, to produce a well-defined cross section, the electron and positron results are just too similar to be usefully plotted.

6. Another interesting observation looking at the blue curve of Fig. 3 is that the RC effect in the entire Q2 range of this work is relatively small and constant, 7-8%, which is interestingly the same

size of the normalization systematic uncertainty quoted in this paper. In some way, this makes a direct comparison with the e+p scattering results very interesting.

The overall size of the correction is driven by the relatively large acceptance of the radiative tail because of the limited momentum resolution compared to high-precision spectrometers. A large part of it comes directly from QED calculations, is charge-sign independent and is believed to have a very small uncertainty. Indeed, a smaller cut would have likely a larger correction, but an even smaller uncertainty. The numerical similarity to our normalization uncertainty is purely accidental.

Answer to Referee B:

The paper reports a first measurement of the charge-averaged elastic scattering cross section of electrons/positrons on protons by the OLYMPUS collaboration.

Compared to either electron-proton or positron-proton measurements separately, the charge-averaged differential cross section is far less sensitive to hard two-photon-exchange (TPE) effects, conferring some advantage in the extraction of the proton form factors in the measured $Q^{\wedge 2}$ range.

High-level summary and recommendation: the measurements reported in this paper represent an ancillary result of the OLYMPUS experiment, and they appear to be valid and publishable. They would have a reasonably significant impact on the determination of the proton form factors in the $Q^{\wedge 2}$ range from 0.6-2 $GeV^{\wedge 2}$, a range which is already well-covered by previous experiments, but in which the discrepancy between extractions of the proton's electric/magnetic form factor ratio from cross sections and polarization observables starts to become significant. However, I cannot recommend the paper for publication in PRL without substantial revisions, and unless the issues limiting the impact of the data are addressed, as described below:

The OLYMPUS detector was optimized for a measurement of the ratio of the e+p/e-p cross sections, and the results were already published in PRL in 2017 (Ref. [26] of the manuscript). As such, the systems for monitoring of the luminosity were optimized for the determination of the relative luminosity between e+ and e-.

On the other hand, OLYMPUS was not designed for precise absolute cross section measurements or monitoring of the absolute luminosity, and the

large global normalization uncertainty quoted in the paper reflects this.

Given the unambiguous theoretical expectation that hard TPE contributions to the cross section are suppressed in the charge-averaged cross section, these new data can in principle reduce the theoretical uncertainties in the extractions of the proton form factors from lepton-proton elastic scattering cross sections in the measured $Q^{\wedge 2}$ range.

However, the large global normalization uncertainty of the absolute cross sections makes the impact of these data somewhat less clear and convincing. In principle, only the relative $Q^{\wedge 2}$ dependence of the charge-averaged cross section is determined precisely, whereas the quoted uncertainty in its absolute global normalization encompasses a significant fraction of the entire range of the data and fits shown in Fig. 2. In other words, one can move all the data points up or down together by 7.5%. If the global normalization uncertainty is really 7.5%, I think this significantly weakens the conclusions of the paper and limits the additional insight into the proton form factors that can be gained from these new data.

We agree with the referee that a better absolute normalization would be helpful, and in ideal circumstances, we would certainly have wished for one. However, we believe that the impact of the missing normalization is very limited: At the lowest Q^2 points of the new OLYMPUS data, existing extractions agree to within maybe 2 percent. An absolute normalization of 2% for data not connected to small Q^2, where the absolute value of the formfactor at Q^2=0 can help with the determination, is already at the level typically quoted for dedicated high-precision, absolute measurements. Fixing the normalization of the OLYMPUS data to a chosen fit, or the average of several, would then propagate this uncertainty to the higher Q^2 points.

In the case one compares the new OLYMPUS data to fits, assuming the OLYMPUS data would have a smaller absolute normalization uncertainty of, say, even only 1%, but would otherwise be unchanged, one would still use the same procedure to compare fits to the data, and would not really learn any new information, since the required normalization shifts are still all acceptable.

In a global fit with floating normalization that includes the Mainz data, the normalization of data sets with any overlap to the Mainz data is essentially completely fixed to the per-mille level even without any constraints on the normalization from the data sets themselves, i.e. such a fit only profits from the shape information of the added data, not from its normalization.

Of course, an absolute normalization on the sub-percent level would have an impact, alas, this would have been quite a different experiment, and in fact, is currently probably outside of our capabilities at the given kinematics.

On the other hand, I might argue for quoting a smaller global normalization uncertainty based on the fact that two independent methods for the determination of the absolute luminosity (the "slow control" and "SYMB" methods) were found to agree at the 1% level, with each in principle providing a "cross check" of the other. How probable is it that two independent measurements of the same quantity, one with an uncertainty of 7%, and one with an unknown uncertainty, would agree at the 1% level? Suppose, for example, that the "slow control" method had an uncertainty of 5%. Then the difference between the two methods would have a standard uncertainty of about 8.6%. The probability that the two methods would then agree with each other at the 1% level by chance would only be about 5%. So I might argue for quoting a significantly smaller global normalization uncertainty, although if it is really true that the uncertainty of the "slow control" method could not be reliably quantified, then this would be a statistically dubious proposition, and accidentally good agreement between two independent measurements would not, on its own, constitute sufficient evidence to quote a smaller global uncertainty.

We had essentially the same argument internally when we were putting on the finishing touches on the paper. However, we concluded that the agreement between the two systems—each with potentially large biases—only tells that the biases of the two systems are similar, not that the biases are small.

But even if we would believe in, say a 2% total luminosity uncertainty, our total normalization error would be only half of what we have now. This, in turn, is still large compared to ≈2% uncertainty in the form factors at the lowest Q2 points, i.e., the impact on a global fit would still come primarily from the shape, rather than the normalization.

While we regret that better confidence in the normalization couldn't be achieved, since our honest accounting doesn't permit a smaller uncertainty, and the impact of such an improved normalization would be limited anyway, we stand by the 7%.

Moreover, while the advantage of the suppression of hard-TPE contributions in this observable is clear and unambiguous, the discussion surrounding Fig. 3 is far less convincing. Unless I have missed something, the fact that the charge-odd radiative corrections other than hard-TPE become a significant fraction of the total as $Q^{\wedge 2}$ increases seems to be of limited relevance to the systematics of form factor extraction, UNLESS the theoretical uncertainties/model-dependence of the charge-odd RC other than hard-TPE are significantly larger than the theoretical uncertainties

of the charge-even ones. The authors don't present any evidence or argument as to why this should be the case. The authors claim that because all the charge-odd RC are suppressed, this makes the cross section less sensitive to uncertainties in the RC prescription. But they don't quantify how much of an advantage this gives in comparison to other sources of systematic uncertainty. While the large uncertainties surrounding the hard-TPE contribution are well established since no model-independent prescription for such corrections exists, I think that if the authors want to go further and claim some benefit from suppressing the charge-odd contributions to the standard RC, then they should be quantitative as to how much of a benefit there is in terms of systematic uncertainty reduction. Otherwise, I think that the advantage is clear enough in terms of the suppression of hard-TPE contributions without resorting to the vague and ambiguous discussion surrounding Fig. 3, which I would advocate be removed unless the surrounding discussion is made quantitative.

The referee is in principle right, of course. If all of the correction uncertainty stems from the even part, this would not reduce the total uncertainty. If one distributes the uncertainty just proportionally to the size of the correction, the impact is small. But we would argue that this is not the case. Indeed, it seems that most difficulty comes from the description of radiative processes off the proton. These effects tend to be small, while the relatively larger corrections from the lepton part are better in control.

Looking at charge-odd vs charge-even effects, proton-side diagrams and lepton-side diagrams mix for the odd, making the total contribution larger than the proton-side with proton-side combinations for the charge-even corrections. We therefore would argue that indeed the cancelation can provide fundamental benefits in the reduction of uncertainty. We think it's worthwhile to investigate this further, but we feel that a full analysis is beyond the scope of this paper and would require a substantial campaign together with theorists. There is a lot of interest in this field, as highlighted for example by https://arxiv.org/abs/2012.09970

To try to make this argument more clear, we have reformulated the paragraph:
"The advantage of charge-averaging technique is that it suppresses all of the charge-odd radiative corrections. The suppression of hard TPE is advantageous because of the uncertainties associated with calculating it, but there may be additional benefits as well. Bremsstrahlung from the proton poses a similar challenge to hard TPE since it depends on an off-shell proton current. The interference term between electron- and proton-bremsstrahlung is one of the suppressed charge-odd effects, which, combined, grow in magnitude to become a sizable fraction of the total correction at higher Q^2, shown in Fig. 3. By forming the charge average, the dominant part of the radiative correction is from radiation from the electron legs, which is under better theoretical control."

In addition, I have a few minor comments and suggestions for where the

clarity of the paper could be improved:

Page 1, last paragraph: "cross section, which are" --> "cross sections, which are" (subject-verb agreement)

We agree and changed the text accordingly.

Page 2, third full paragraph: "The left-right symmetry... was used as a cross-check in the analysis" is vague. Cross-check of what? Maybe this sentence could either be expanded and clarified or removed. Or perhaps a reference to the previous OLYMPUS papers would be sufficient.

Since this is covered on p. 3, last full paragraph, we opt to remove it here.

Page 2, bottom left: unclear what is meant by the statement that the magnetic field was mapped "in-situ".

We replaced in-situ with: "*with the magnet unmoved from its final position in the experiment.*"

Also: saying the Monte Carlo is "sophisticated" is quite subjective. Perhaps better to say "detailed" or "realistic"?

We replaced it with realistic. We note that we also added a sentence explaining the including of time-dependence.

Page 2, right column: The paragraph on "Track reconstruction used a fast hierarchical... two distinct track fit algorithms" begs for elaboration. Why are two track fit algorithms needed if one works sufficiently well? What do you gain from the comparison? Perhaps providing a reference here for the interested reader would be Appropriate.

We added a few sentences to explain:

"*The design of the drift chambers and the running conditions in OLYMPUS led to some track-fitting ambiguities that were difficult for the algorithms to resolve. While the algorithms did well for most constellations, they failed for certain pathological cases. However, the two algorithms struggled in different cases, so that the combination of both algorithms secured the reconstruction with high efficiency over the whole phase space.*"

Systematic uncertainty discussion, page 2, bottom: it's not entirely clear what is meant by a "possible" difference between "simulated elastic efficiency" (which is not clearly defined) and "that of the experimental data". It seems like you could both simulate the efficiency and independently determine it from the data. If so, then either there was or wasn't a difference. These two sentences could be made more clear.

We reformulated the sentence to make it clearer:

"Within the precision of the study (≈1%), there was no indication of inefficiency beyond that caused by ToF and drift chamber inefficiencies, and we therefore assign a 1% normalization uncertainty for any tracking inefficiency. An additional 2% absolute normalization was estimated for other sources not tested by this method, e.g. for the trigger efficiency."

Page 3, left: is it really necessary to state that you propagated the uncertainty associated with the background subtraction to the final results? We assume you know how to propagate errors.

We have eliminated it.

I was somewhat surprised that a "screen for optimal running conditions" removes approximately 1/3 of the total integrated luminosity. How does this compare to any similar screen applied in the e+p/e-p cross section ratio analysis published in PRL 2017?

This is mainly driven by two facts: We did not use the data taken in the first period at all, because of a target malfunction. Further, we rigorously eliminated any runs with trips in the ToF HV or beam hiccups, DAQ errors etc. This screen is the same as for the 2017 PRL, and we modified the sentence to read:

"The total recorded data were screened for optimal running conditions, and a subset corresponding to 3.1 fb^{-1} of integrated luminosity (the same subset as in [27], 2017 PRL) was selected for the results presented here."

Page 3, right: the paragraph discussing the separation of the point-to-point and normalization uncertainties is a bit unclear and confusing. The four independent analyses as far as I can tell differ only in their elastic event selection routines. You say that you "minimize the difference to the average", in allowing the normalization of each of the four analyses to vary. But "minimizing the difference" of what to the average of what? The point-by-point averages? Is the average itself allowed to float? Because varying the normalization of each data set will also lead to varying the average,

so the average is itself a moving target. Anyway, I think I understand what you did and seems superficially valid, but this discussion could be made much more clear.

Yes, we first build the average, and then split the variance into a normalization-scatter around the average, and a point-to-point scatter after the normalization. We believe this text makes it clearer:

"*We report the cross section determined from the average of the results of the four independent analyses. We further use the variance between the analyses to estimate systematic uncertainties from event selection choices. However, we first remove the effect of normalization differences between the analyses. We find, for each analysis, the normalization factor that minimizes the difference of the analysis to the original average. After renormalization, we then determine the remaining variance and use this as an additional point-to-point uncertainty. The std. deviation of the normalization constants, 1.5\%, is added as an additional contribution to the global normalization uncertainty.*"

Prompted by the comment, we also noticed a couple of bugs in our averaging code, which have been fixed, with small changes in the table.