We would like to first thank the editor and reviewers for acknowledging our work and all the positive comments. Please find our responses below.

Reviewer #1: Summary of my comments

This is an excellent paper. It is really well written and structured. It provides a nice literature review and clearly situates the paper's contributions against previous studies. I find the method proposed rather convincing and robust and I believe the paper makes a great contribution to the field of transport geography and accessibility. Having said that, I only have a few minor comments and suggestions that I believe the authors need to address. Please see my comments below.

Minor comments:

1. p.8 "Tribby & Zandbergen (2012) improved transit accessibility by incorporating detailed representations of the sidewalk network for traveling to, from and between transit stops." It's important to recognize the advance of Tribby & Zandbergen's paper in 2012. Nonetheless, the authors seem to dismiss some recent but important advances in the literature. Since 2012, accessibility modeling has become significantly more detailed with the development of OpenTripPlanner (OTP), even though OTP also assumes static transit headways. More recently, though, the new R5 routing engine uses Monte Carlo simulations to generate many possible timetables from a static GTFS feed in order to simulate uncertainties in headways over a time window. This allows one to generate estimates of travel time uncertainties even with static GTFS feeds. This method is well documented in this paper (Conway, Byrd, & van Eggermond, 2018) and the application is illustrated in this other paper (Pereira, Saraiva, Herszenhut, Braga, & Conway, 2021).

We thank the reviewer for the suggestion. We add suggested papers in the last paragraph in section 2.2 and paragraph “uncertainties” in section 2.3. Please find the inserted red fonts in the updated draft.

2. Figures 6 to 9. I understand the measurement unit of the schedule-based accessibility unreliability is in percentage points. Correct? If so, I'd suggest make this clear in the Y axis of these figures. I would also suggest the authors could explain a bit better in the text what does it mean to have a value of 0.6 or 1.2? What is the substantial interpretation of these percentages?

Yes, the measurement is a percentage. We change the mentioned graphs’ Y axis from numeric to percentages per your suggestion. We also add a short paragraph in section 4.2 in the front of the result part to briefly revisit the definition and scale of the measure. Please find the red font part in section 4.2 or below.

“The percentage can be interpreted as the scheduled-based measure’s deviation from the realizable measure (see equation (4)), therefore representing the inaccuracy or overestimation of the scheduled-based measure: for example, 0% means no overestimation at all and the scheduled-based measure is the same as the realizable measure; while 100% means the scheduled-based measure is twice as much as the realizable measure.”

3. Figure 8. I would suggest labelling the x axis with text (Monday, Tuesday, …) instead of numbers.

Thank you. Please find the updated graph.

4. Figure 5. It looks like that the correlation between a stop unreliability and its closeness centrality becomes higher when considering higher travel times. This is a simple observation of something that perhaps deserves attention in future research. No need to respond to this.

Yes, this is an important result. Figure 6 also shows the same phenomenon.

5. Short provocation here playing the devil's advocate. One could say that the proposed realizable accessibility is of little use to transit users because it only captures accessibility levels at an instant for a specific moment in time. Other passengers will not get to experience that accessibility level again since access conditions are constantly and continuously changing.

This is an important comment. We agree that each result can only represent a specific instant moment. However, empirical insights require more than only one moment. We gave an example of practical applications of the proposed method in Figure 7 – 9. We calculate the realizable accessibility and the schedule-based unreliability for each day in two years, different day of the week, and different hours. Some recurring patterns can be found in these results, which are a combination of pattens in many instant moments; and it can guide future planning/operation in both strategic and refined ways.

Meanwhile, each instant moment can reveal many results that a non-specific measure cannot show. With the proposed method, a transit authority, scientist, or a transit planning app can calculate realizable accessibility at any given specific time and location, as the tool has the ability. In that sense, we would like to argue that a time/location specific measure is instead a major advantage since it shows a more refined pattern, as we argued in the literature review. This can be regarded as an example of the future trend of “more refined measure”. However, we do acknowledge that this will take longer time and more computational power.

6. Is the algorithm used by the authors publicly available? If so, I would recommend including a link to the code repository or adding the code scripts as supplementary material to paper on the journal website. This will help others use the method proposed in the paper, make this a stronger contribution to the literature.

Yes, we are willing to make the code open source. Please find the links in the supplementary materials or here: <https://github.com/luyuliu/COTA-AccessibilityReliability>.

7. p.8 "However, their analysis assumes travel \*though\* planar space…" Typo.

Thank you, we fix the typo. Please find the updated word in red font.

Reviewer #2: Realizable Accessibility: Evaluating the Reliability of Public Transit Accessibility using High-resolution Real-time Data

It is an excellent work on accessibility in public transport using real-time data. The paper proposes a new indicator of accessibility in public transport (real-time accessibility), considering delays in bus systems and imperfect user information on systems operations. This also proposes the concept of accessibility unreliability to measure overestimation of schedule-based and retrospective accessibility measures. Big data is used for calculations.

The paper makes a relevant contribution to the literature. It is very well structured and very well written.

In the opinion of this reviewer, it can be published as is.

Thank you so much for the kind words!

Reviewer #3: The study and the results are fascinating but suggest some changes:

The article is very critical of the GTFS dataset that it presents not present actual data, etc. It should be remembered how many works were based on this data format. And the level of underestimation of travel times is in places where there are no bus lanes and is problematic mainly when it comes to buses, which was also pointed out by the authors.

This is a great insight. We acknowledge that GTFS static is a very useful data source, and it definitely helps to establish many well accepted works. We are still actively using it for different purposes as GTFS static is very well formatted and convenient when it comes to accessibility models. However, the overestimation (of accessibility) or underestimation (of unreliability) is a valid systematic issue like we discussed in the paper. It is not realistic to advocate that future accessibility works should totally abandon GTFS static for good, considering the lack of accessible real-time data/technologies right now; instead, we have several suggestions:

* Static data are still useful. But we should acknowledge the systematic risk of overestimation when conducting accessibility analyses with static data.
* Similar to the suggestion of reviewer 1, we can use simulations to incorporate unreliability in the modelling of accessibility even with static data.
* We should develop infrastructure for real-time data and prepare to transition from static data to real-time data. This includes more transparent GTFS real-time API, data collection, and computational platform.

We have similar arguments in the second last paragraph in the conclusion section.