Please find our responses to the valuable comments.  
  
  
Reviewer #1: This paper employs the space-time prism method and GTFS data to analyze the impacts of long-term and short-term disruptions on the Central Ohio Transit Authority (COTA) bus system in Columbus, Ohio. The study defines accessibility unreliability as the deviation between realizable accessibility and scheduled accessibility, providing a measure to assess the reliability of accessibility. The conclusions drawn from the analysis hold practical implications for local transportation planning and management. However, considering the publication guidelines set forth by JTG, this paper is not recommended for publication in the journal due to the following reasons:

We thank the reviewer for the comments. We are glad to hear that the reviewer acknowledges the practical implications of our research for local transportation planning and management. We are also relieved to know that the reason why the reviewer did not recommend the paper for publication is not because of the paper’s quality.

However, we do not agree that this paper’s contribution is limited to the local context. The generalizability of the case studies is twofold: 1) The methodologies introduced in the paper are very generalizable and not specific to this geographic context. 2) On the other hand, the disruptive events introduced in the paper, such as football games and the COVID-19 pandemic, are also not unique to Columbus. College football games are the most popular and most disruptive sport events in the United States, and we are still experiencing the impacts of COVID-19 in a global level. We firmly believe that this paper is within the scope of Journal of Transport Geography.

Insufficient Literature Review: The literature review lacks comprehensiveness and exhibits signs of selective inclusion. The three research gaps emphasized by the authors have already been addressed in multiple related publications in the field of complex networks and public transit planning. Consequently, the claimed contributions of this paper are difficult to substantiate convincingly.

We thank the reviewer for the comment. Please find our revision to the literature review below in our responses to comment #1 of reviewer 2.

On the other hand, despite being a classic topic in network science and transportation engineering, the prior studies lack a spatiotemporal perspective, and most studies do not address the unique characteristic of public transit networks (i.e., time dependent, discrete, and schedule-based). Prior studies also do not address the problem with empirical real-time data.

Ambiguity in Definitions: The definitions of "Short-term disruption" and "Long-term disruption" are presented multiple times in the paper, leading to confusion. While it is relatively straightforward to comprehend disruptions based on their temporal duration, issues arise when disruptions fall within a time span of 2-6 days. Because the paper clarifies that the short-term disruptions are short in time span: typically, not exceeding a single day, while the long-term disruptions are longer in time span, which can last from weeks to months and years.

We thank the reviews for the comments. This is a good observation. We agree that it can be hard to categorize when it comes to disruptions of 2 – 6 days. Besides the duration of the event, another major distinction between short-term and long-term is whether the event would impact the schedule of the transit system (introduced in line 39, pp 4 and line 7, pp 5), and if the system would recover after the event cease to exist (introduced in line 8, pp 5).

To reflect the comment, we change *weeks to months and years* to *several days to months and years.* Please check line 7, pp 5 for your reference.

Methodological Considerations: While the authors emphasized the application of GTFS data is a valuable aspect of this research and has led to conclusions not previously explored, it is worth noting that the use of widely adopted public transit card data, with comparable precision and granularity, could have been an alternative approach.

We thank the reviewer for the comment. We agree that smart card data is an equally important data source for public transit research. However, there are several drawbacks that make it hard to compete with GTFS real-time data for the research questions we are trying to answer in this paper:

* First, smart card contains only the information about the actual ridership, while GTFS real-time data contain all buses and potential trips information. This is crucial and a prerequisite for understanding the real-time performance of the transit system.
* Second, smart card data are hard to access, while GTFS real-time data are open data and very accessible in many systems.
* Finally and most importantly, many transit systems, if not most transit systems in the United States, do not have smart card data at all, as these systems do not have a smart card system. This includes the city of Columbus and COTA system.

Superficial Conclusions: The findings of the study appear intuitive and do not surpass the scope of existing research or practical knowledge. Although the paper uncovers certain phenomena, it fails to delve into the underlying mechanisms behind the decline in accessibility and reliability. The lack of investigation into whether accessibility and reliability reduction are caused by high passenger volume at certain stations or station failures limits the paper's depth. The implications drawn from the conclusions might be relevant to local transportation planning in Ohio but cannot be generalized to other regions.

This is a good point, and this can be considered as a limitation of this study and the dataset. We cannot do causality analysis due to the empirical nature of the dataset, and we chose not to delve into the correlation analysis between reliability and ridership, as the paper would significantly exceed the recommended word limits.

With that all being said, however, we do not agree that this limitation would diminish the contribution of this paper entirely for three reasons.

1. As the review admitted earlier, the empirical findings and methods presented by this paper are indeed not previously explored.
2. With the methods introduced in the paper, any cities with GTFS-real-time feed can replicate the system and detect the impacts of disruptions in any geographic and temporal contexts.
3. Meanwhile, the empirical studies in the paper are also useful for other cities and transit systems, as other cities also have football games and experience the COVID-19 pandemic.

In conclusion, the reviewer suggests that this paper may be more suitable for submission to a journal that primarily focuses on case study articles rather than JTG.

Reviewer #2: This is an interesting manuscript that has good potential to be published in the JTG. However, I recommend a major revision of the manuscript because I argue that there are many rooms to be improved.

We thank the reviewer for the useful comments. Please find our detailed response below.  
  
#1. Literature review: It would be more beneficial to discuss more papers on the topic. For example, there are some empirical studies (e.g., Beck et al., 2020; Kim & Kwan, 2021) that report more "resiliency" of human mobility with respect to the pandemic. Although the author(s) 's paper largely focuses on transit, it would be fruitful to discuss how overall human mobility has changed related to the pandemic.  
  
Beck, M. J., Hensher, D. A., & Wei, E. (2020). Slowly coming out of COVID-19 restrictions in Australia: Implications for working from home and commuting trips by car and public transport. Journal of Transport Geography, 88, 102846.  
Kim, J., & Kwan, M. P. (2021). The impact of the COVID-19 pandemic on people's mobility: A longitudinal study of the US from March to September of 2020. Journal of Transport Geography, 93, 103039.

Thank you for the recommendation. We added these two papers to our review and more discussions. We also made major revision to the literature review section (please see section 2.3, pp 5) to reflect your and other reviewers’ comments.   
  
#2. Section 3.2: There are already well-developed transit-based shortest travel time calculation tools, such as r5r (Pereira et al., 2021), but why do the authors need to develop a time-dependent Dijkstra algorithm? Also, if the authors develop the algorithm, how do the authors validate its results? In other words, how are the authors confident about the shortest travel time result obtained from their own algorithm?  
  
Pereira, R. H., Saraiva, M., Herszenhut, D., Braga, C. K. V., & Conway, M. W. (2021). r5r: rapid realistic routing on multimodal transport networks with r 5 in r. Findings.

This is a great question. The reason why we developed a new routing engine is because other traditional engines, such as the legendary r5r, do not support real-time calculation and lack functions to simulate people’s route choice and calculate their travel time under the guidance of real-time trip planning app. Therefore, r5r cannot calculate the reliability of accessibility. This is a major methodological contribution of this paper, proceeded by Liu et al. (2022).

In terms of the issue of validation, our time-dependent Dijkstra algorithm is based on many well-tested papers in network science (please check pp 6 for more information), and the FIFO assumption is empirically tested. Since we do not know any other packages that can do the same, we cannot conduct similar performance tests conducted in Pereira et al. (2021). On the other hand, without behavioral survey or GPS data, we really cannot validate if the shortest travel path calculated in a routing engine can indeed represent an actual experience and accessibility of a user, which has been proven to overestimate people’s mobility in many prior studies (Liu et al., 2022; Wessel et al., 2017; Wessel & Farber, 2019). Therefore, this can indeed be a limitation for this study as well as all accessibility studies that used routing engines.

#3. Section 3.2. "One vehicle overtaking another in violation of the FIFO restriction is a rare event: we estimate from COTA data that 95% of the buses meet this restriction." How do the authors control the "95%" level in their shortest travel time algorithm?

We did not control the level, but we empirically calculated the rate of buses that meet the restriction. The value of the rate happened to be 95%.

#4. Section 3: Please consider providing a brief section that provides the context of the study area, such as maps illustrating population density, major activity locations (e.g., football stadium, downtown), and major transportation networks. Since this crucial information is not presented, it was difficult to understand the geographic context of the study area. For instance, future readers of the paper should know that this analysis is based on Columbus (Ohio), which is a large-sized metro area, so their findings and implications might not be readily applicable to many other college towns, such as Lafayette (Purdue), Urbana-Champaign (UIUC), Charlottesville (U of Virginia), etc.

This is a great suggestion. Please find the added section in line 35, pp 5. We also added a map with pop density, bus routes with frequency, and locations of downtown and football stadium in Figure 1.   
  
#5. Results: I encourage the authors to improve the graphic quality of the figures. For instance, Figure 1 seems to be generated without careful editing in graphics, such as alignment, margin, color, font, etc. This is a minor issue and does not affect the scientific quality of the manuscript, but, considering the high impact of the journal (JTG), the figures can be improved.

Thank you for the suggestion. We updated the figures.

#6. Results: Please consider providing tables that report the descriptive statistic of key values, such as accessibility scores each day. This would be helpful to understand the results that are only presented in figures.

#7. Results (Section 4.1): This can be an important question related to the study design. How are the authors confident that the short-term disruption is only affected by the football game rather than other issues, such as weather (e.g., thunderstorms), accidents, and others? Were the authors able to control these external factors so that they could purely observe the impacts of the football game? At least the authors may want to provide detailed available contexts (e.g., whether there were serious traffic accidents or major weather events, such as thunderstorms and tornado warnings) and discuss the potential limitations of not controlling those factors. Moreover, regarding the research design, the authors can consider including more dates that do not host games and comparing the results. The current research design does not allow us to understand the true impact of football games on accessibility scores. Lastly, many universities, on game day, provide additional or special transit services to address unusually-high travel demand. Did the authors consider this or at least discuss it (i.e., potential impacts on the analytical results)?

First, as stated in the limitation, although we are very confident about the short-term disruption is affected by the football game, we do not have enough data to prove the causality between the high unreliability and the disruptive events (stated in the final paragraph, first point).

Second, we are confident because:

1. We collected the weather data, including temperature and precipitation, and found no significant anomalies in the record.
2. Similarly, there is no major accidents happening in those game days.

In figure 4 (originally figure 3), we selected home game, away game, and non-game days and compare. We found that unreliability in home game days is significantly higher than non-game days. This strongly suggests that services in game days are more reliable because of the football games, not other factors. This is very similar to the analysis recommended by the reviewer.

Finally, we do not consider other university transit services in our analysis, as the scope of the paper is public transit services provided by the COTA system. In the other words, the focus of all the analyses is the COTA transit system, not the game days. A potential impact could be that the additional service may remedy some negative impacts on COTA buses; however, this falls out of the scope of the paper, and we lack empirical data to address the problem.

#8. Figure 2: It is unclear how to interpret Figure 2. For example, what does it mean by peak? Please consider providing clear descriptions of this figure.

#9. Figure 3: Given the fact that game hours are different on each date (see Figure 1), how could this aggregated figure tell meaningful information? It might be helpful to have an x-axis in a different way, such as an hour from the game start time. For instance, 0 denotes the game start hour, -1 indicates 1 hour before the game start hour, and +1 indicates 1 hour after the game start hour. By doing so, the figure can consider different start hours of games.

This is a great observation. We chose to use the aggregated average figures because the aggregated figures show that unreliability on home game days is systematically higher than away game days and non-game days, and away game days are systematically higher than non-game days. The empirical distribution of the curved to be averaged does not concern the physical meaning of the aggregated hourly profile. In other words, it may not be able to maintain the shape of the curves, but it can indeed show the average level for each curve and compare each curve’s values for each hour, which is the whole point of the analysis.

#10. Figure 6: The legend (color break) can be misleading because the red indicates negative values in the first figure but not in the second figure. Please fix this cartographic issue.

This is a good question. We chose to use opposite color scheme because of cartographic considerations. Based on cartographic practice and common sense, red suggests warning and worse situation, while blue suggests normal and better situations. Please notice that the second graph shows unreliability change, and a negative unreliability value indicates that the service becomes better. Therefore, red color in both maps indicates worse performance.

#11. (Page 15) "The two measures' spatial patterns are very similar:": This argument can be assessed more quantitatively. At least, the authors can consider producing kernel density maps or providing a summary table (e.g., accessibility score change average in terms of downtown vs. suburban, etc.) to provide evidence for their claim. From the perspective of the reviewers, this figure does not support the argument.

#12. (Page 15) "We find that football games are correlated with exceptional high unreliability in local public transit system." Could the authors consider revising this phrase? It sounds a bit weird to say that "football games are correlated" with something. Probably, a better way to say can be that football games affect the unreliability of the local public transit system.

This is a good point. We change the sentence to “We find that **the presence of football games** is correlated with exceptional high unreliability in local public transit system”. We would like to be cautious and avoid claim about the causality between football games and unreliability.

#13. (Page 16) "We suggest that more public transit systems should use real-time data to monitor system performance and guide future system." This statement seems to be less relevant to the manuscript. Many transit agencies already use real-time data (e.g., real-time dashboards). The implications might focus on the methodological limitations of existing studies that do not utilize real-time data. However, more fundamentally, this study did not show the conventional approach (that does not use real-time data) is misleading the finding related to the disrupted transit services and accessibility scores. Thus, I argue that this implication cannot be derived based on this study's findings. Similarly, the policy implications of this study could be more specific. It is trivial to say that transit agencies should address uncertainties caused by events. Probably, the authors can consider highlighting the spatial pattern of their findings (e.g., Figures 4 and 6) so that their findings can provide important implications for formulating place-based transit policies.

#14. Limitation (Page 16): Another limitation can be the potential MAUP issue with the analysis. For instance, the station-level analysis might not be useful for formulating the policy (especially equity-related analysis), and thus, aggregated-level analysis can be preferred in some cases (Javanmard et al., 2023).  
  
Javanmard, R., Lee, J., Kim, J., Liu, L., & Diab, E. (2023). The impacts of the modifiable areal unit problem (MAUP) on social equity analysis of public transit reliability. Journal of Transport Geography, 106, 103500.  
  
We added the limitation to the final part. Thank you for the recommendation.  
  
Reviewer #3: This study examined the resilience of transit system, by comparing the scheduled and actual travel time of COTA in Columbus, Ohio under two scenarios: the football game and the COVID-19 pandemic. The topic is interesting and the methods are solid. I just have some minor suggestions:  
  
1. I would suggest the authors better justify how COVID-19 pandemic might influence transit reliance. It would definitely affect transit ridership, but since the study focused on accessibility reliance, it would be nice to discuss how it might impact on the accessibility and travel time.  
  
2. It is a bit hard to recognize the spatial pattern from the figures. Besides the variance across stations, maybe the authors could try to plot the variation across raster grids or census tracts, to better visualize the pattern.