Annenberg Institute
BROWN UNIVERSITY

EdWorkingPaper No. 25-1239

Understanding Disruptions: Causes of and Variation in Lost Instructional Time

Xander Beberman
Stanford University

Sarah Novicoff
Public Policy Institute of California

Ana Trindade Ribeiro
Stanford University

Carly D. Robinson
Stanford University

Susanna Loeb
Stanford University

Virtual instruction has boomed after the COVID-19 pandemic, including the use of virtual environments within in-person schools. But, research has provided little evidence about student experiences on these virtual platforms, nor how to improve the use of these platforms. Through natural language processing techniques, this study examines over 26,000 virtual tutoring sessions that took place within in-person schools to identify the frequency of disruptions during virtual learning. We find that 81 percent of allocated instructional time was undisrupted. Technology problems consumed approximately nine percent of planned instructional time, while student disruptions consumed another seven percent. We find substantial variation across schools, with fewer disruptions when tutoring occurred in dedicated environments such as libraries or tutoring-specific classrooms.

VERSION: October 2025

Suggested citation: Beberman, Xander, Sarah Novicoff, Ana Trindade Ribeiro, Carly D. Robinson, and Susanna Loeb. (2025). Understanding Disruptions: Causes of and Variation in Lost Instructional Time. (EdWorkingPaper: 25-1239). Retrieved from Annenberg Institute at Brown University: https://doi.org/10.26300/k6da-t861

Abstract_ Virtual instruction has boomed after the COVID-19 pandemic, including the use of virtual environments within in-person schools. But, research has provided little evidence about student experiences on these virtual platforms, nor how to improve the use of these platforms. Through natural language processing techniques, this study examines over 26,000 virtual tutoring sessions that took place within in-person schools to identify the frequency of disruptions during virtual learning. We find that 81 percent of allocated instructional time was undisrupted. Technology problems consumed approximately nine percent of planned instructional time, while student disruptions consumed another seven percent. We find substantial variation across schools, with fewer disruptions when tutoring occurred in dedicated environments such as libraries or tutoring-specific classrooms.

Virtual instruction is increasingly common. More than half a million K-12 students attended fully virtual schools in the United States in 2021-22 (Miron et al., 2023). Even for students who attend K-12 schools in-person in traditional school buildings, some portion of instruction increasingly takes place through virtual environments. Quantifying the scale of educational technology usage in school is not straightforward, but one report from a company that partners with hundreds of school districts to monitor technology usage suggests that the average K-12 student in the U.S. spends about 100 minutes a day on school-issued devices in a classroom (Bennett, 2025). Students commonly type essays in Google Docs with in-person teacher supervision, play a quiz game with in-person classmates to review course content on Kahoot!, or participate in virtual tutoring.

Given an unusually strong research base that provides evidence that tutoring can be one of the most effective educational interventions for struggling students (Nickow et al., 2024), and amidst substantial learning loss after the COVID-19 pandemic (Kuhfeld et al., 2022), schools

have expanded their provision of tutoring. A 2024 federal survey found that 78 percent of public schools provided tutoring during the 2024-25 school year. Among those who did not offer tutoring, many expressed interest but noted barriers to pursuing tutoring; in particular, 36 percent of respondents identified staffing challenges as an obstacle (*School Pulse Panel: Surveying High-Priority, Education-Related Topics*, 2024).

Virtual tutoring is one solution to staffing challenges because virtual programs can leverage broader labor markets instead of being constrained to individuals who can commute to a particular school site. Districts are increasingly taking advantage of virtual tutoring. In Ector County, Texas, the district spent $5 million for virtual tutors to work with 6,000 students during the school day, while Wichita Public Schools in Kansas spent $2.5 million on virtual tutoring, and Mississippi signed an $11 million contract for virtual tutoring statewide (Belsha, 2022).

Despite the growing prevalence of these virtual environments, we know relatively little about students' experiences in these settings. Studies have shown the potential of virtual tutoring to provide substantial benefits for students (Carlana & La Ferrara, 2024; Deacon & Chojnacki, 2023; Gortazar et al., 2023; Roschelle et al., 2020), but not all studies of virtual programs have found strong positive effects (Ready et al., 2024; Torgerson et al., 2016). The variation in program effects may be due to differences in program quality, differences in the population served, or differences in the strength of the comparison groups' experiences, but these differences in effectiveness may also be driven by elements of the environment itself that limit or facilitate students' experiences.

This study presents among the first pieces of evidence about the virtual tutoring experience (Abdelshiheed et al., 2024; Demszky et al., 2024; Zhang et al., 2025). Here we focus specifically on disruptions during tutoring sessions. Virtual tutoring sessions may be prone to

disruptions from the technology itself, which may or may not work, and students may or may not know how to use. Disruptions can also come from background noise, from other students in the classroom, or from a range of other factors. This research sheds light on the extent and heterogeneity of disruptions, as well as the relationship between disruptions and the implementation context. In doing so, we identify effective approaches for implementing virtual learning environments to minimize disruptions.

The study also offers insight into disruption measurement. Researchers have traditionally tallied disruptions from the back of the room or surveyed teachers, methods that can be both costly and imprecise. Instead, this study leverages virtual tutoring transcripts and natural language processing (NLP) tools to measure how and when disruptions occur at substantially lower cost, less time and greater accuracy. We quantify potential learning time at scale across 26,778 tutoring sessions conducted in spring 2023.

We find that, on average, 81 percent of tutoring session time is undisrupted by technology, background noise, or students. Disruptions tend to take up 19 percent of a tutoring session, with technology-based disruptions consuming 9 percent of the time and student-caused disruptions consuming 7 percent. These disruptions vary significantly by age group, suggesting that virtual tutoring may be a more promising intervention for slightly older students, in this case second graders versus kindergarteners. Disruptions also vary by tutoring location, suggesting that schools can reduce disruptions by moving tutoring into different spaces with dedicated staffing (e.g., libraries or tutoring-specific classrooms).

Background

Instructional Time Literature

Learning takes time. Prior meta-analyses of the effect of time in school have shown that losses in instructional time can reduce achievement and that increases in instructional time can increase student achievement (Gromada & Shewbridge, 2016; Holland et al., 2015; Kraft & Novicoff, 2024; Patall et al., 2010). But, allocated instructional time (i.e., the amount of time in which a school system intends instruction to take place) varies widely across the United States — the top 10 percent of schools in total instructional time offer almost 200 hours more in a year than schools in the bottom 10 percent (Kraft & Novicoff, 2024). Not all time that is allocated to instruction is necessarily spent learning, though. External interruptions (Kraft & Monti-Nussbaum, 2021), teacher absences (Herrmann & Rockoff, 2012), student absences (Goodman, 2014; J. Liu & Loeb, 2021), and even students' own disruptive activities can erode the total time in which students could be learning.

Natural Language Processing (NLP) in Education

When researchers have previously tried to measure total time loss, they have historically dispatched individuals to sit in the back of a room and manually track when disruptions occur in a ledger (Leonard, 1999). This method is both time-consuming and costly, as well as subject to potential biases, as students and teachers potentially adjust their behaviors due to the presence of an outsider in the classroom. Those same studies have often complemented observational data with teacher or administrator surveys, but those rely on perhaps imperfect memory of disruptions that are often less than a minute in length but can collectively add up to a lot of time. NLP techniques offer a new opportunity to measure these disruptions more accurately and at lower cost, and thus better estimate potential learning time.

NLP methods use computational power to interpret human language at scale, typically by training a specialized model to identify the phenomenon of interest in the language presented.

Though NLP has been around for decades, the last decade has seen a boon in education research using these techniques, thanks in part to advances in their accuracy and ease of access (Xue, 2024). For example, in Markowitz et al. (2023), researchers used 3.5 million office discipline records to identify language patterns in the way teachers described student behavior by race. Similar methods have been used to analyze essays written by pre-service teachers to understand their reasons for wanting to join the profession (Bartanen et al., 2024). Reading millions of office records or thousands of essays by hand would have otherwise taken years to draw these same conclusions; in this way, NLP has both sped up analyses while also allowing them to be conducted at a greater scale and in a greater array of settings than previously possible with traditional qualitative research.

Both those studies used NLP to analyze typed text, but NLP can also be paired with high-quality transcription to understand what occurs in live settings and document more complex elements of the educational experience. For example, Tan & Demszky (2023) used NLP methods to identify patterns of teacher speech during classroom management components of a lesson and find punitive language occurs with more frequency in classrooms with high proportions of Black students. In another example, Zhang et al. (2025) used NLP tools to identify who educators paid attention to in group tutoring sessions and found attention disparities across achievement levels and demographic lines, particularly by gender. Across these two studies and many others, researchers have employed NLP methodology to analyze interactions between educators and students to understand disparities in our education system.

Contribution and Goals of the Current Study

The current study extends the use of NLP methods in educational settings to understand instructional time provision in virtual tutoring. We aim to answer the following research questions:

1) How often do disruptions occur in virtual tutoring? What causes these disruptions, and how many are attributable to virtual learning specifically rather than early elementary learning environments generally?
2) How do disruptions vary by lesson group size, grade, timing of a lesson, and student demographics?
3) How do disruptions vary by school site? Controlling for student demographics, what characteristics of a school site's tutoring program predict disruptions?

Data

We use a large dataset of audio and video recordings of virtual tutoring sessions originally collected for a randomized controlled trial evaluating the effectiveness of an early literacy tutoring program in an individualized (1:1 student-tutor) or paired (2:1 student-tutor) format (Robinson et al., 2024). These tutoring sessions took place across 12 elementary schools in a charter network in Texas. Our data includes 26,778 sessions that took place between November 2022 and May 2023, of which 19,448 were conducted 1:1 and 7,150 were conducted 2:1. An additional 180 sessions were conducted with a ratio greater than 2:1 due to tutor absences. We received separate recordings for the tutor and student(s) present in each session, with metadata that includes student and tutor identifiers and their timestamps for logging in and off of each session.

We observed 1,357 students and 192 tutors participating in virtual tutoring during this time. Tutoring took place with students in kindergarten, first, and second grade, and focused on

learning literacy skills. The number of students and tutoring contexts varied across the campuses. On three campuses, students were tutored in their classrooms while other learning was happening around them, whereas on seven campuses, students were tutored outside of the classroom in the library, cafeteria, or a designated tutoring classroom. On the remaining three campuses, the location of tutoring varied by grade.

We linked school administrative data from the charter network to tutoring sessions. Specifically, the administrative data we use includes variables such as student grade, date of birth, race/ethnicity, gender, whether the student received free or reduced-price lunch or was otherwise indicated as economically disadvantaged based on the receipt of other public assistance, whether they had an Individualized Education Plan or 504 Plan, whether they were designated as an English learner, and their availability for tutoring within the school day. It also included student scores on two previously validated literacy assessments: Dynamic Indicators of Basic Literacy Skills (DIBELS) and MAP Reading. We supplement these data with information about the implementation setting in each school provided by the tutoring provider and school enrollment numbers provided publicly by the Texas Education Agency.

Methods

This study uses tutor transcripts from 26,778 sessions described above to identify disruptions, classify their cause, and identify associations between disruptions and student and school characteristics.

We focus exclusively on tutor's speech for our analysis due to significant differences in the quality of recordings for students and tutors in the same session, which impacted the accuracy of the transcripts. While qualitatively exploring student and tutor audio, we noted that tutors' recordings are very understandable: tutors are generally connected to tutoring sessions

from an environment with almost no background noise, and tutors tend to speak clearly into the microphone. In contrast, student recordings from the same tutoring sessions contained a wide range of background noise levels in the student environment and highly variable student speech clarity. In a review of transcriptions of 11 student audio recordings, we found that automatic speech recognition tools consistently performed poorly on student audio. Models frequently hallucinated nonexistent or incorrect speech, confused background speakers with tutees, or missed large segments of quiet or noisy audio. Automated transcripts of tutor speech, however, were reliably accurate. As a result, a limitation of this study is that we were unable to use student transcripts in combination with the tutor transcripts to determine student-tutor speech turns.

Our use of tutor transcripts allows us to document every time a tutor pauses a session to address a disruption in the learning environment. This approach introduces a limitation in that we do not count disruptions too brief or insignificant to be noted by the tutor, or disruptions that a tutor chooses to ignore. These omitted disruptions are likely to be small, but their omission does imply our count of disruptions likely underestimates the true magnitude of disruption in a tutoring session.

Our first step in quantifying disruption was to transcribe tutor audio in each session using the automated speech transcription tool WhisperX (Bain et al., 2023). We annotated and classified disruptions at the utterance level, where each utterance is a discrete chunk of speech output by WhisperX. In general, each utterance corresponds to about one sentence or standalone phrase.

Second, we defined what counts as a disruption and what categories of disruptions should be in our analysis. To do so, we randomly selected 18 sessions to read both transcripts and listen to their session audio. These sessions were chosen to be representative within audio quality (e.g.,

clear, ok, and unclear) and group size. Three members of the research team then worked together to distinguish common types of disruptions that can occur in the context of in-school, individualized, or paired virtual tutoring. We then conducted a larger pilot phase in which the same three members of the research team annotated a sample of about 255 transcripts to validate that consistent ratings between annotators was feasible within this framework and that additional categories did not need to be added. We randomly selected five utterances each from transcripts of 255 tutoring sessions, with the constraint that at least one utterance had to come from the first two minutes of the tutor and student appearing in the session, as, anecdotally, this is when tutors reported that many disruptions occurred.

To complete annotations at the pilot stage and also at our larger annotation stage, we used Label Studio (Tkachenko et al., 2020), an online labeling interface to annotate utterances from tutor transcripts. In the labeling interface, we displayed the selected utterances along with the three utterances before and after to give context. Annotators were able to select one of the six categories (seen below) for each utterance and also indicate when a line in the context around the utterance showed a different disruption. Disruptions in the context lines were later annotated using the same system.

This pilot stage validated a disruption framework that included the following categories: background noise, second student, student disruption, substitute tutor, and technology problems, as well as a category for other stop to usable instructional time. Descriptions of the categories with examples as given to our annotators are shown in Table A1.

Third, after our framework was validated by this pilot stage, we expanded our pool to include 24 annotators who were provided with a copy of the framework and a 20-minute presentation describing the project goals. All 24 annotators completed annotations, but inter-rater

reliability of the full group was substantially lower than that of a 22-annotator subset, as shown in Table A2. Therefore, we restricted our training data to the 2,501 annotations of 1,488 unique utterances completed by the more reliable 22 annotators. The frequencies of each disruption category in this training dataset are shown in Table A3. For these annotations, we observe a Krippendorff's alpha of 0.71 when considering all disruption classes, and of 0.74 when considering only whether or not an utterance was a disruption, which suggests moderate inter-rater reliability.

As the fourth step, we fine-tuned a text classifier using our training dataset to identify types of disruption typical to virtual tutoring sessions and annotate tutor transcriptions at scale. We used the RoBERTa classifier (Y. Liu et al., 2019), a pre-trained transformer-based large language model that takes in text and predicts whether it is one of multiple user-determined classes, as a base. This classifier allows us to identify when a disruption occurs during a tutoring session and estimate the time spent on each type of disruption based on session transcript timestamps. We used 80 percent of our 2,501 annotations as the training set, and the remaining 20 percent of observations were held out for use in a test set. The resulting model achieved an accuracy of 80 percent on the test set. An alternative binary version of our classifier that does not consider the disruption category, only whether or not the utterance indicated a disruption, resulted in higher accuracy, 85 percent. The biggest challenge in developing our multi-label classifier was that two categories, substitute tutor and other stop to usable time, had very few occurrences in our data, limiting our ability to reliably evaluate our model's performance on these categories. We found that excluding these categories from our analyses reduces the estimated amount of total disruption by less than one percentage point and does not significantly impact other results. For this reason, we used the complete model, including predictions for those

categories, to report our results. Further evaluation metrics for the disruption classifier, as well as additional details on the training and test set, can be found in Appendix B. We then used the fine-tuned model to classify all tutor utterances in our dataset.

Finally, in order to estimate the time associated with each disruption in a session, we use the average of the predicted probability of each disruption category for all utterances in a session, weighted by the length of time between the start of the utterance and the following utterance. We estimated the following regression model for each category of disruption as a separate dependent variables. Our unit of observation is the student-session level, which attempts to describe the disruption that an individual student experiences in a single session. We include an indicator for 2:1 tutoring that equals 1 for sessions conducted with a preassigned pair and 0 otherwise; a vector of student characteristics indicators including economically disadvantaged status, English learner designation, student with disability designation, recorded race Black, recorded ethnicity Latina/o/e, and recorded gender female. We also include beginning-of-year DIBELS score in three terms: a standardized beginning-of-year score, the standardized beginning-of-year score squared to recognize that the relationship between achievement and disruption may not be linear, and a binary variable for whether a student earned the minimum possible score for their grade level to proxy for students who may struggle more broadly than a generically low score represents. Finally, we include school-specific indicators (s); and grade-specific (g) indicators. We use these estimates to explore how variables representing student characteristics and how each school in our sample relates to the different types of disruption. We cluster standard errors at the student and session level.

$$y_i = \beta_0 + \beta_1(2:1t)_i + \mathbf{X}_i\boldsymbol{\gamma} + \delta_{s(i)} + \lambda_{g(i)} + \epsilon_i$$

We estimate a second group of regressions to investigate the relationship between implementation characteristics and disruptions, also at the student-session level. We included the following independent variables in addition to student controls: where and at what time of day tutoring took place in the school, number of students being tutored simultaneously at the school, and the share of simultaneous sessions over total students enrolled as a proxy for strain on infrastructure capacity. In our regressions, disruptions and share of simultaneous sessions are encoded as percentage points for ease of interpretation.

Findings

RQ1: How often do disruptions occur in virtual tutoring? What causes these disruptions, and how many are attributable to virtual learning specifically rather than early elementary learning environments generally?

Disruptions overall consumed approximately 19 percent of allocated instructional time in the virtual tutoring sessions. The most common source of these disruptions was the technology itself, whether Internet outages or non-functional microphones or screens, among other challenges (e.g., “You can’t see me? I’m not sure why you can’t see me” or “My thing keeps glitching”). These technological disruptions consumed an average of nine percent of allocated instructional time in a session, more even than disruptions caused by early elementary students who are unfocused on the lesson at-hand and force the instruction to pause (e.g., “There’s no need for writing right now so just leave your pencil alone.”) Those student disruptions, the second most common cause of disruption, consume seven percent of total time. Other common causes of disruption include background noise (e.g., “Sorry, did you say something? It was hard to hear.”) and the addition of a second student whose late entry into a lesson creates a need for

the instructor to pause (e.g., “So, [student name], I want to let you know what I already told [second student name]”). Figure 1 Panel A visualizes these disruptions.

RQ2: How do disruptions vary by lesson group size, grade, timing of a lesson, and student demographics?

Disruptions occur more frequently in 2:1 sessions than in 1:1 sessions. The allocated instructional time of 2:1 sessions was typically eroded by 24 percent compared to the allocated instructional time of 1:1 sessions, which was typically eroded by 18 percent instead (a difference of 6 percentage points or 25 percent). This difference is statistically significant ($p < 0.01$) and is visualized in Figure 1 Panel B. The data show a statistically significant increase in technology problems, potentially caused by the entrance of a third device into the session, and also by a statistically significant increase in time the tutor spends catching up a second student who enters the session late.

Disruptions also occur with greater frequency for younger children. Students in kindergarten experienced sessions that were 20.1 percent disrupted compared to students in first grade who experienced 19.6 percent disruption and in second grade who experienced 18.7 percent disruption. This difference is statistically significant ($p < 0.01$) and is visualized in Figure 1 Panel C. Though student behavior itself, which generally becomes less disruptive over time throughout an elementary school trajectory, is a component of these disruption differences, even when that is excluded, we still see statistically significant differences in disruption rates by grade. Older students may be better equipped to troubleshoot technological challenges, thus reducing the amount of time these challenges detract from a session.

Disruptions also vary over time. They are concentrated at the beginning of a session. As a lesson flows, the frequency of disruptions_especially technology-based disruptions_declines

steadily as visualized in Figure 1 Panel D. Meanwhile, Figure 1 Panel E shows that disruptions are also more frequent at certain times of the year, most notably spiking after winter break. That feature of disruptions is not driven by technology but rather by student-driven disruptions that have been previously documented in in-person environments (Darling-Hammond et al., 2023).

Finally, the frequency of disruptions varies by student characteristics. Table 1 provides these results for disruptions overall and for each disruption type. Female students, as well as Latina/o students and students classified as English learners experienced fewer disruptions, controlling for differences in group size, grade level and initial test performance. We do not find statistically significant differences by economic disadvantage or by classification for special education, though economically disadvantaged students do experience statistically more disruptions due to technology and students with disabilities do experience more disruptions due to students and to background noise.

RQ3: How do disruptions vary by school site? Controlling for student demographics, what characteristics of a school site's tutoring program predict disruptions?

The degree to which disruptions erode allocated instructional time varies by school site, as seen in Figure 2 Panel A. Students on the most disrupted elementary school campus lost 23 percent of their allocated instructional time to disruptions while students on the least disrupted elementary school campus lost only 15 percent of their allocated instructional time (a statistically significant difference of eight percentage points). The difference remains after controlling for student demographics as shown in Figure 3. The magnitude of this difference is comparable to the difference in disruption between 2:1 and 1:1 sessions and larger than the difference in disruption across grades.

Controlling for student demographics and group size, program features predict the degree of disruption that students experience in their tutoring sessions. As seen in Table 2 Panel A, tutoring that occurs in-class rotations (i.e., when some students in a classroom are in virtual tutoring sessions while other students are working with an in-person instructor) is the most disrupted. In contrast, tutoring sessions that take place in school libraries experience almost seven percentage points less disruption, while tutoring that takes place in tutoring-specific classrooms (i.e., classrooms where all present students are working with a tutor virtually on an Internet-connected device) experience 4.2 percentage points less total disruption time.

Table 2 Panel B shows that differences by location are driven primarily by technology problems. This may be because educators in tutoring-specific spaces are more accustomed to and able to help remedy technological problems when they arise. For example, when a kindergarten student encounters a technical problem in a classroom with only one teacher who is also working with other students, that student is likely responsible for troubleshooting themselves. On the other hand, if that same kindergarten student were completing tutoring in a library or a classroom where all other students were on their own devices with educators, then the educator in the room would be able to help troubleshoot any technical problems that arose.

In addition to the space in which tutoring takes place, we include additional terms in the regression for the number of simultaneous tutoring sessions, both in absolute terms and as a percentage of the school enrollment, to control for the school size and its likely physical and technical capacity. We estimate a coefficient of -0.082 for variations in the number of simultaneous sessions and 0.810 for the associated change as a share of enrollment (Table 2 Panel A), which suggests that sessions in smaller schools are more sensitive to the number of simultaneous sessions compared to larger schools. For example, when a school with 500 students

adds 5 simultaneous sessions, which increases the ratio of simultaneous sessions to enrollment by one percentage point, that is associated with an increase of 0.4 percentage points in disruption (-0.082*5 + 0.810 = 0.4). The equivalent increase in the ratio for larger schools is associated with a negligible increase or even lower levels of disruption during tutoring sessions, though. This suggests that small schools should be aware of their capacity—both the capacity of their staff and of their technology, as seen in Table 2 Panel B—to manage simultaneous sessions when scheduling, whereas larger schools may not be as constrained by these factors.

Lastly, we include indicators for the time of day when sessions occurred, grouping sessions into three periods: early morning, midday, and late afternoon. Compared to mid-day sessions, the reference group, Table 2 Panel A shows that early morning sessions experience lower levels of disruption, which is partly explained by fewer tech-related disruptions, shown in Panel B. This difference might be explained by the number or types of tasks school staff engage with in each of these periods as well as changes in student concentration throughout the school day.

Discussion

The typical K-12 student in a traditional school building spends more than an hour and a half of each school day (about 25 percent) online (Bennett, 2025). As virtual education becomes more common, an understanding of virtual environments can help educators improve student experiences and leverage the new opportunities more effectively. In this study, we analyze transcripts of virtual tutoring sessions that took place with elementary students in traditional school buildings specifically to understand the role of disruptions. We used an iterative approach to define learning disruptions in a virtual tutoring environment and then deployed 22 annotators to label approximately 1,500 unique utterances. These annotations were used to fine-tune a

RoBERTa classifier, which we applied to a sample of 26,778 tutoring session transcripts for our analysis.

We find that disruptions consume about 19 percent of allocated instructional time with technology problems accounting for the largest share, nine percent of allocated instructional time or 47 percent of total disruption. Student disruptions accounted for the second largest share, seven percent of allocated instructional time or 37 percent of total disruption. These estimates may be lower bounds because our measure of disruption is based on tutor transcripts. In other words, if a disruption occurred but a tutor did not comment on it, either because it was brief or because the tutor chose to ignore it, then we do not count it.

The data show substantial variation across schools: students at the most disrupted elementary school campus lost 23 percent of their allocated instructional time to disruptions, compared to just 15 percent at the least disrupted campus. This eight-percentage point difference represents a 35 percent reduction in instructional time lost to disruptions. We also find that disruptions occurred more frequently in 2:1 tutoring sessions compared to 1:1 sessions and with greater frequency for younger students.

Taken together, these results suggest that virtual learning environments may more successfully convert allocated instructional time to potential learning time in smaller groups and for older students, who are better equipped to work through the inevitable technological challenges. This research also suggests, however, that the erosion of instructional time by technology and other elements of the virtual classroom is not inevitable and can be successfully mitigated by schools. In particular, this study indicates that schools experience less disruption when tutoring takes place in quieter, tutoring-specific spaces like libraries or dedicated classrooms, where staff may be better able to troubleshoot technical issues. Additionally, we find

evidence that increasing the number of simultaneous sessions can strain the capacity of a school's technological infrastructure, particularly at small school sites, if not adequately planned for.

As virtual learning becomes more commonplace in school, this study sheds light on a practical, yet underexplored, question: how much instructional time are students actually receiving in these virtual environments? By quantifying learning time lost to disruption and identifying strategies to reduce those disruptions, this research offers actionable insights for schools aiming to maximize the value of virtual instruction. Beyond the practical implications, the study also pioneers a new methodological approach to measuring disruptions—moving beyond traditional in-person observation to leverage NLP in a way that is unobtrusive and scalable. This approach not only enhances the precision of measuring how time is spent in virtual environments, but also opens new avenues for research into instructional interactions and quality across both virtual and in-person settings. As educational technology continues to evolve, these tools and investigations will be essential for designing effective learning environments.

References

Abdelshiheed, M., K. Jacobs, J., & K. D'Mello, S. (2024). Aligning Tutor Discourse Supporting Rigorous Thinking with Tutee Content Mastery for Predicting Math Achievement. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial Intelligence in Education* (pp. 150–164). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-64299-9_11

Bain, M., Huh, J., Han, T., & Zisserman, A. (2023). WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *INTERSPEECH 2023*, 4489–4493. https://doi.org/10.21437/Interspeech.2023-78

Bartanen, B., Kwok, A., Avitabile, A., & Kim, B. H. (2024). Why Do You Want to Be a Teacher? A Natural Language Processing Approach. *Educational Researcher*. https://doi.org/10.3102/0013189X241276814

Belsha, K. (2022, June 29). Schools are spending millions on new virtual tutoring. Is it working? *Chalkbeat*. https://www.chalkbeat.org/2022/6/29/23186973/virtual-tutoring-schools-covid-relief-money

Bennett, A. (2025). *Breaking Down Student Screen Time | Lightspeed Systems*. Lightspeed Systems. https://www.lightspeedsystems.com/blog/breaking-down-k-12-screen-time-what-every-parent-and-school-leader-should-know/

Carlana, M., & La Ferrara, E. (2024). *Apart but Connected: Online Tutoring, Cognitive Outcomes, and Soft Skills* (Working Paper No. 32272). National Bureau of Economic Research. https://doi.org/10.3386/w32272

Deacon, G., & Chojnacki, G. (2023). Impacts of UPchieve On-Demand Tutoring on Students'

Math Knowledge and Perceptions. Middle Years Math Grantee Report Series. In *Mathematica*. Mathematica. https://eric.ed.gov/?id=ED628646

Demszky, D., Wang, R., Geraghty, S., & Yu, C. (2024). Does Feedback on Talk Time Increase Student Engagement? Evidence from a Randomized Controlled Trial on a Math Tutoring Platform. *Proceedings of the 14th Learning Analytics and Knowledge Conference*, 632–644. https://doi.org/10.1145/3636555.3636924

Goodman, J. (2014). *Flaking Out: Student Absences and Snow Days as Disruptions of Instructional Time* (No. w20221; p. w20221). National Bureau of Economic Research. https://doi.org/10.3386/w20221

Gortazar, L., Hupkau, C., & Roldan, A. (2023). Online tutoring works: Experimental evidence from a program with vulnerable children. In *EdWorkingPapers.com*. Annenberg Institute at Brown University. https://www.edworkingpapers.com/ai23-743

Gromada, A., & Shewbridge, C. (2016). Student Learning Time: A Literature Review. *OECD Education Working Papers*, Article 127. https://ideas.repec.org//p/oec/eduaab/127-en.html

Herrmann, M. A., & Rockoff, J. E. (2012). Worker Absence and Productivity: Evidence from Teaching. *Journal of Labor Economics*, *30*(4), 749–782.

Holland, P., Alfaro, P., & Evans, D. K. (2015). *Extending the School Day in Latin America and the Caribbean*. The World Bank. https://doi.org/10.1596/1813-9450-7309

Kraft, M. A., & Monti-Nussbaum, M. (2021). The Big Problem With Little Interruptions to Classroom Learning. *AERA Open*, *7*, 23328584211028856. https://doi.org/10.1177/23328584211028856

Kraft, M. A., & Novicoff, S. (2024). Time in School: A Conceptual Framework, Synthesis of the

Causal Research, and Empirical Exploration. *American Educational Research Journal*, *61*(4), 724–766. https://doi.org/10.3102/00028312241251857

Kuhfeld, M., Soland, J., & Lewis, K. (2022). Test Score Patterns Across Three COVID-19-Impacted School Years. *Educational Researcher*, *51*(7), 500–506. https://doi.org/10.3102/0013189X221109178

Leonard, L. J. (1999). Towards Maximizing Instructional Time: The Nature and Extent of Externally-Imposed Classroom Interruptions. *Journal of School Leadership*, *9*(5), 454–474. https://doi.org/10.1177/105268469900900506

Liu, J., & Loeb, S. (2021). Engaging Teachers: Measuring the Impact of Teachers on Student Attendance in Secondary School. *Journal of Human Resources*, *56*(2), 343–379. https://doi.org/10.3368/jhr.56.2.1216-8430R3

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach* (No. arXiv:1907.11692). arXiv. https://doi.org/10.48550/arXiv.1907.11692

Markowitz, D. M., Kittelman, A., Girvan, E. J., Santiago-Rosario, M. R., & McIntosh, K. (2023). Taking Note of Our Biases: How Language Patterns Reveal Bias Underlying the Use of Office Discipline Referrals in Exclusionary Discipline. *Educational Researcher*, *52*(9), 525–534. https://doi.org/10.3102/0013189X231189444

Miron, G., Hagle, S., & Gulosino, C. (2023). *Full-Time Virtual Schools: Enrollment, Student Characteristics, and Performance*. National Education Policy Center. https://nepc.colorado.edu/publication/virtual-schools-annual-2023

Nickow, A., Oreopoulos, P., & Quan, V. (2024). The Promise of Tutoring for PreK–12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence. *American*

Educational Research Journal, *61*(1), 74–107. https://doi.org/10.3102/00028312231208687

Patall, E. A., Cooper, H., & Allen, A. B. (2010). Extending the School Day or School Year: A Systematic Review of Research (1985–2009). *Review of Educational Research*, *80*(3), 401–436. https://doi.org/10.3102/0034654310377086

Ready, D. D., McCormick, S. G., & Shmoys, R. J. (2024). The Effects of In-School Virtual Tutoring on Student Reading Development: Evidence from a Short-Cycle Randomized Controlled Trial. In *EdWorkingPapers.com*. Annenberg Institute at Brown University. https://edworkingpapers.com/ai24-942

Robinson, C. D., Pollard, C., Novicoff, S., White, S., & Loeb, S. (2024). The Effects of Virtual Tutoring on Young Readers: Results from a Randomized Controlled Trial. In *EdWorkingPapers.com*. Annenberg Institute at Brown University. https://edworkingpapers.com/ai24-955

Roschelle, J., Cheng, B. H., Hodkowski, N., Neisler, J., & Haldar, L. (2020). *Evaluation of an online tutoring program in elementary mathematics*. Digital Promise. https://files.eric.ed.gov/fulltext/ED604743.pdf

School Pulse Panel: Surveying high-priority, education-related topics. (2024, October). National Center for Education Statistics. https://public.tableau.com/views/2025_SPP_CollegeCareerReadiness/OfferAdvancedCourses?:embed=y&:showAppBanner=false&:toolbar=no&:display_count=n&:showVizHome=n&:origin=viz_share_link&:tabs=no&:device=desktop

Tan, M., & Demszky, D. (2023). *Sit Down Now: How Teachers' Language Reveals the Dynamics of Classroom Management Practices*. EdWorking Papers.

https://doi.org/10.26300/9YJ6-JN52

Tkachenko, M., Malyuk, M., Holmanyuk, A., & Liubimov, N. (2020). *Label Studio: Data labeling software*. https://github.com/HumanSignal/label-studio

Torgerson, C., Ainsworth, H., Buckley, H., Hampden-Thompson, G., Hewitt, C., Humphry, D., Jefferson, L., Mitchell, N., & Torgerson, D. (2016). *Affordable online maths tuition: Evaluation report and executive summary*. Education Endowment Foundation. https://durham-repository.worktribe.com/output/1606609

Xue, H. (2024). Development History and Frontier Trends of Natural Language Processing Technology in Education: A Knowledge Graph-Based Visualization Analysis. *Proceedings of the 15th International Conference on Education Technology and Computers*, 436–443. https://doi.org/10.1145/3629296.3629366

Zhang, Q., Wang, R. E., Ribeiro, A. T., Demszky, D., & Loeb, S. (2025). *Educator Attention: How computational tools can systematically identify the distribution of a key resource for students*. EdWorking Papers. https://doi.org/10.26300/HVQ6-D385

Tables and Figures

Table 1: Student-Level Predictors of Disruption

	Tech Problem	Student Disruption	Other Stop To Usable Time	Second Student	Background Noise	Substitute Tutor	Disruption
2:1 tutoring	1.984***	2.510***	0.093***	1.181***	0.545***	0.066***	6.380***
	(0.107)	(0.071)	(0.019)	(0.025)	(0.029)	(0.009)	(0.149)
Economically disadvantaged	0.534***	-0.020	0.027	-0.018	-0.103	0.025*	0.446
	(0.193)	(0.134)	(0.035)	(0.043)	(0.063)	(0.013)	(0.276)
English learner	-0.206	-0.314***	-0.005	-0.012	-0.027	-0.022*	-0.586***
	(0.129)	(0.081)	(0.024)	(0.028)	(0.033)	(0.012)	(0.177)
SWD	-0.086	0.484***	-0.082**	0.038	0.151**	-0.039**	0.467
	(0.237)	(0.182)	(0.040)	(0.057)	(0.069)	(0.017)	(0.361)
Black	0.204	-0.623***	-0.093**	0.118**	0.276***	0.018	-0.100
	(0.238)	(0.155)	(0.041)	(0.046)	(0.075)	(0.014)	(0.325)
Latina/o/x	-0.167	-0.919***	-0.058	0.043	0.219***	0.020	-0.861***
	(0.230)	(0.145)	(0.039)	(0.044)	(0.069)	(0.019)	(0.310)
Female	-0.279***	-0.411***	0.033*	0.092***	0.078***	0.002	-0.485***
	(0.107)	(0.069)	(0.019)	(0.023)	(0.029)	(0.010)	(0.148)
Kindergarten	0.234	0.387***	-0.107***	0.058*	-0.251***	0.030***	0.352*
	(0.152)	(0.096)	(0.023)	(0.034)	(0.039)	(0.009)	(0.208)
Second Grade	-0.661***	-0.392***	0.134***	-0.062**	0.026	-0.010	-0.965***
	(0.122)	(0.080)	(0.024)	(0.026)	(0.035)	(0.012)	(0.170)
Dibels BOY	0.174**	0.006	0.116***	-0.003	0.112***	-0.001	0.404***
	(0.074)	(0.049)	(0.014)	(0.017)	(0.020)	(0.007)	(0.103)
Dibels BOY squared	-0.142***	-0.138***	-0.033***	0.006	-0.074***	-0.001	-0.383***
	(0.044)	(0.026)	(0.008)	(0.009)	(0.012)	(0.003)	(0.060)
Dibels BOY minimum	0.480*	-0.392**	-0.139***	-0.192***	0.093	0.012	-0.138
	(0.275)	(0.161)	(0.031)	(0.048)	(0.057)	(0.020)	(0.364)
School fixed effects	X	X	X	X	X	X	X
Observations	30824	30824	30824	30824	30824	30824	30824
R^2	0.046	0.054	0.012	0.082	0.031	0.032	0.088

Note: *p<0.1; **p<0.05; ***p<0.01. Sample includes all student-session level data between November 2022 and May 2023, excluding records with missing values for either economic disadvantage indicator or student with disability indicator.

Table 2: Relationship between Program Design Features and Disruption

Panel A: Overall Disruption

VARIABLES	(1) Disruption	(2) Disruption	(3) Disruption	(4) Disruption
Tutored in cafeteria	-4.562***	-4.298***	-3.509***	-3.510***
	(0.612)	(0.637)	(1.074)	(1.071)
Tutored in library	-3.411***	-3.317***	-5.120***	-6.629***
	(0.516)	(0.521)	(0.935)	(1.595)
Tutored in tutoring-specific classroom	0.0444	-0.0467	-2.469**	-4.213**
	(0.550)	(0.567)	(1.145)	(1.753)
Simultaneous tutoring sessions			-0.0722	-0.0819*
			(0.0452)	(0.0436)
Sessions / enrollment			0.661*	0.810**
			(0.399)	(0.390)
Started 7-10 AM				-3.239*
				(1.863)
Started 2-5 PM				-1.678
				(1.347)
Observations	30,824	30,824	30,824	30,824
R-squared	0.019	0.025	0.035	0.036
Student covariates		X	X	X
School fixed effects			X	X

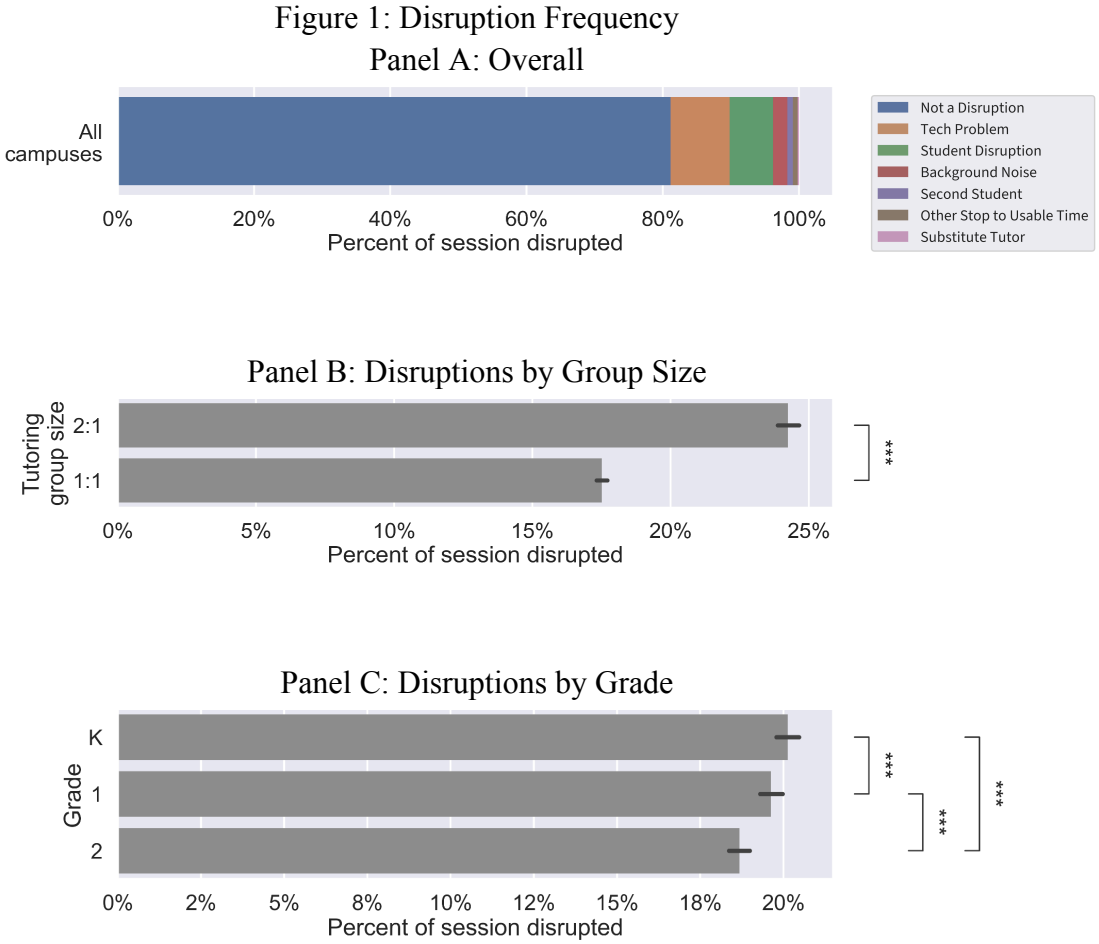
*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. Tutoring space uses in-class rotations as reference category. Tutoring time uses 10 AM-2 PM as reference category. Student covariates include grade, gender, race, economic disadvantage, EL, SWD, beginning of year DIBELS. Simultaneous tutoring sessions calculated as number of sessions starting per hour at each school.

Panel B: Technology Disruptions

VARIABLES	(1) Tech Problem	(2) Tech Problem	(3) Tech Problem	(4) Tech Problem
Tutored in cafeteria	-2.470***	-2.265***	-1.679***	-1.665***
	(0.344)	(0.356)	(0.594)	(0.591)
Tutored in library	-2.989***	-2.886***	-5.135***	-5.526***
	(0.286)	(0.295)	(0.462)	(0.689)
Tutored in tutoring-specific classroom	0.329	0.366	-1.761***	-2.294***
	(0.323)	(0.338)	(0.608)	(0.809)
Simultaneous tutoring sessions			-0.0546**	-0.0686**
			(0.0271)	(0.0270)
Sessions / enrollment			0.586**	0.701***
			(0.247)	(0.245)
Started 7-10 AM				-1.866*
				(0.963)
Started 2-5 PM				-0.382
				(0.626)
Observations	30,824	30,824	30,824	30,824
R-squared	0.023	0.026	0.037	0.038
Student covariates		X	X	X
School fixed effects			X	X

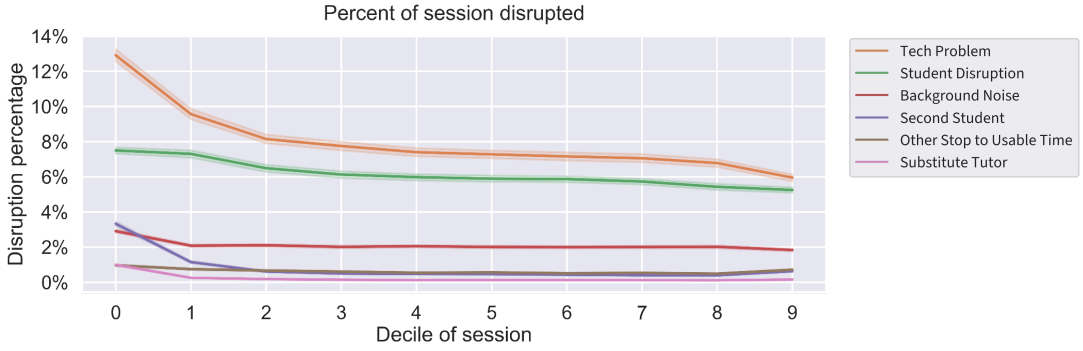
*** p<0.01, ** p<0.05, * p<0.1. Robust standard errors in parentheses. Tutoring space uses in-class rotations as reference category. Tutoring time uses 10 AM-2 PM as reference category. Student covariates include grade, gender, race, economic disadvantage, EL, SWD, beginning of year DIBELS. Simultaneous tutoring sessions calculated as number of sessions starting per hour at each school.

Figure 1: Disruption Frequency



where *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$

Panel D: Disruptions Over the Course of a Session



Panel E: Disruptions Over the Course of a Semester

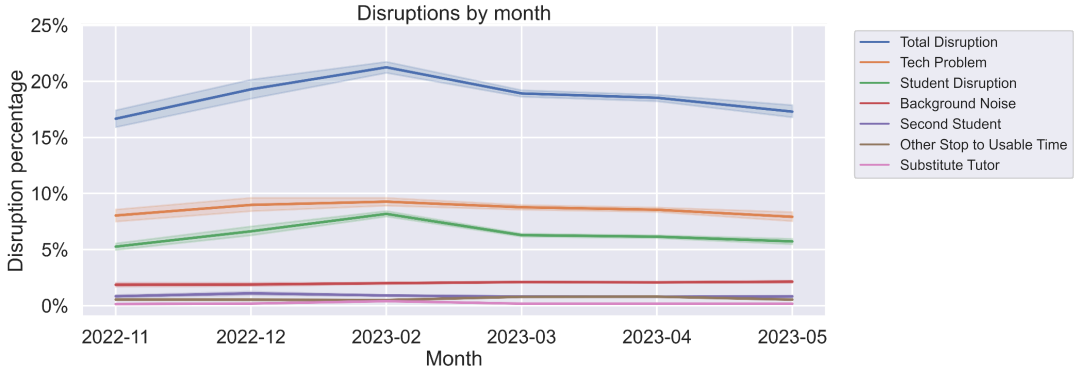


Figure 2: Disruptions by School

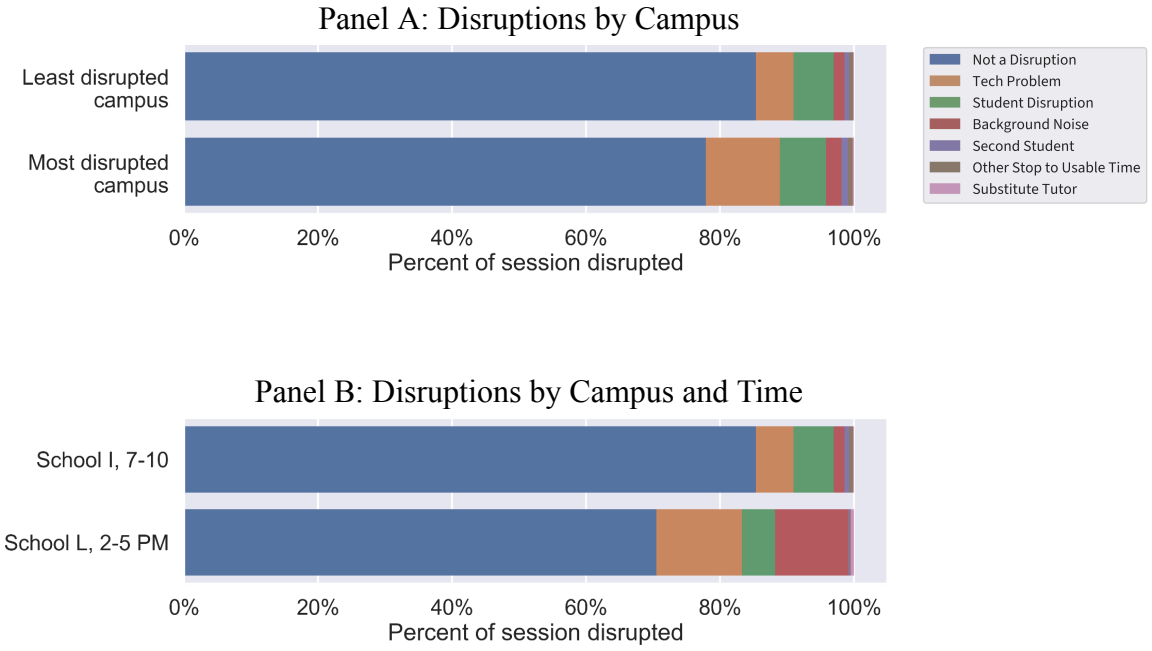
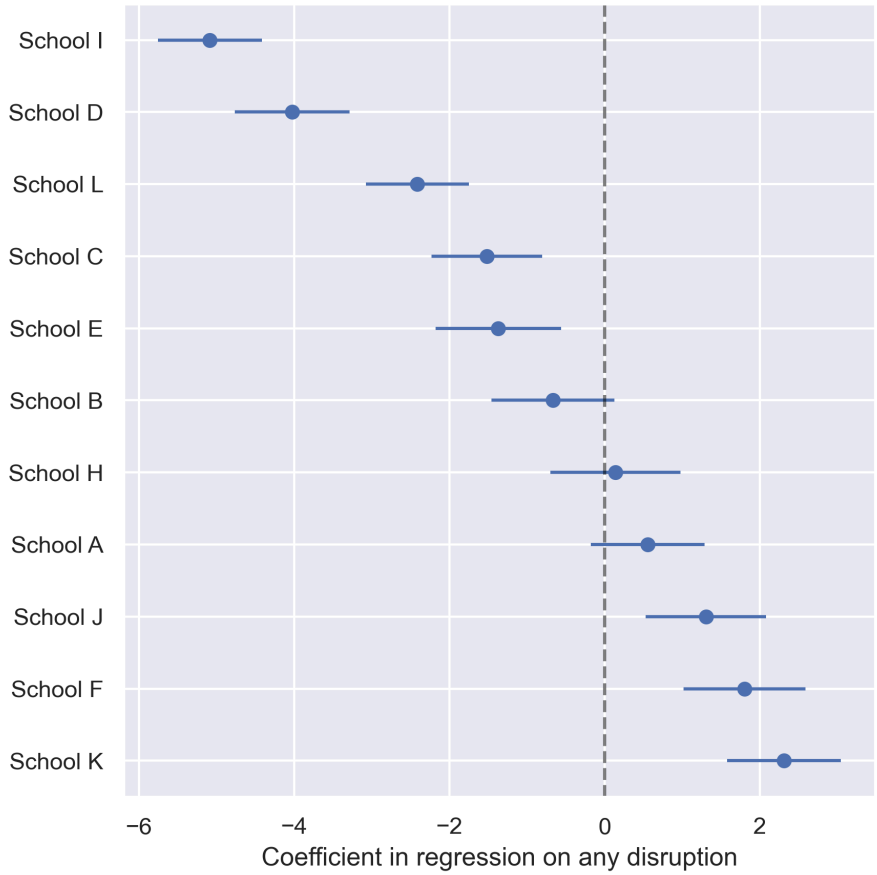


Figure 3: Regression Coefficients for Disruptions by School



Note: School G was chosen as the reference category as it has the median level of disruption per session (19%). Regression is performed at the student-session level and also includes an indicator for 2:1 tutoring, student economically disadvantaged status, English learner designation, student with disability designation, recorded race Black, recorded ethnicity Latina/o/e, recorded gender female, beginning-of-year DIBELS score, and grade.

Table A1: Descriptions of disruption categories given to annotators

Background Noise: The tutor is prevented from delivering instruction because there is too much noise. This includes noise caused by adults or students in the room with the target student. This is often indicated by a tutor asking the student to repeat themselves. Example: “I cannot hear you. It is so loud.”
Second Student: The tutor could be delivering instruction or building rapport during this time period, but they are instead asking or commenting on the absence of a second student or greeting that second student when they arrive. Example: “So [Student A], I want to let you know what I told [Student B].”
Student Disruption: The tutor could be delivering instruction or building rapport during this time period, but the student is doing something that prevents this from occurring. This includes when the student purposefully uses otherwise functional tech to prevent instruction from occurring. Example: “[Student], there’s no writing right now. So just leave your thing. I’m asking a question.”
Substitute Tutor: The tutor could be delivering instruction or building rapport during this time if the regular tutor was present, but the regular tutor is absent and so the substitute tutor has to spend time introducing themselves. Example: “I will be your tutor for today only because your regular tutor is not available today, so maybe she’ll be here on Thursday, all right?”
Tech Problem: The tutor is prevented from delivering instruction because the technology underlying the session (e.g.,headphones, Internet, video, microphone) is not working as intended. This is sometimes indicated by a tutor who does not think they can be heard or seen, or a tutor repeating themselves. Example: “You can’t see me? I’m not sure why you can’t see me.”
Other Stop to Usable Time: The tutor is prevented from delivering instruction or building rapport for unidentified reasons.
Not a disruption: Tutor is delivering instruction or building rapport during this time. Example: “So for our first activity,we are going to be reading the words on the board.”

Table A2: Inter-Rater Reliability for Annotators.

Label Annotators	All categories	Binary disruption
All annotators	0.48	0.54
Subset annotators	0.71	0.74

Note: 24 annotators completed annotation. Our analyses rely on the subset of 22 annotators depicted above whose inter-rater reliability is much higher than the full sample.

Table A3: Frequency of disruption classes of annotated tutor utterances

Disruption category	Count
Not a Disruption	1,948
Tech Problem	239
Student Disruption	140
Background Noise	81
Second Student	63
Other Stop to Usable Time	21
Substitute Tutor	9
Total	2,501

Appendix B: Model Evaluation Process

We used 80 percent of our 2,501 annotations as the training set, and the remaining 20 percent of observations were held out for use in a test set. The training and test sets were split randomly while stratifying on annotated disruption class and keeping annotations of the same utterance constrained to one side of the split. To account for the naturally occurring imbalance in category representation in our corpus, where the majority of utterances is expected to be classified as ”Not a Disruption,” we oversampled disruptive utterances in our training set by a factor of three. The utterances input into the model were also surrounded by context utterances, three before and three after, consistent with the protocol used for human annotation. We used our held out utterances to create a test set of 289 unique utterances with labels agreed on by at least two human annotators.

In addition to producing accuracy metrics discussed in the paper main text, the resulting model also produced the model evaluation metrics seen in Table B1. Due to its very low base rate, no utterances with two agreeing annotations of the category “Substitute Tutor” made it into our final evaluation set, despite being included as a category in our model. The category “Any Disruption” is an aggregate of whether the predicted class of an utterance is any of the disruption categories. Precision (PPV) measures how many of the predicted positives were accurate while recall (sensitivity) measures how many of the positive items were predicted. F1 score is the harmonic mean of precision and recall. ROC-AUC is the area under the receiver operating characteristic curve, which corresponds to the probability the model will score a randomly chosen positive class higher than a randomly chosen negative class. “Any Disruption” is a binary measure constructed from other disruption categories, so it has no predicted probability and therefore no ROC-AUC value.

Table B1: Model evaluation metrics for disruption classifier on the test set.

Category	Precision	Recall	F1 Score	ROC-AUC
Not a Disruption	0.93	0.86	0.89	0.90
Tech Problem	0.60	0.79	0.68	0.94
Second Student	0.75	0.60	0.67	0.97
Student Disruption	0.50	0.72	0.59	0.92
Background Noise	0.35	0.43	0.39	0.74
Other Stop to Usable Time	0.50	0.25	0.33	0.78
Any Disruption	0.67	0.83	0.74	