**Research Questions:**

We will recruit 180 teachers who teach 4th-8th grade math in two cohorts to participate in our study. We will randomly assign them to two coaching conditions–a reflective approach vs. a directive approach. The randomization will be then conducted within coaches, meaning that half of the teachers a given coach works with will be assigned to either  a reflective or  a directive approach. We will answer the following four research questions.

Primary:

1. When integrated with automated feedback, what is the relative effectiveness of a reflective vs. a directive coaching approach on teachers' teaching practices?
2. When integrated with automated feedback, what is the relative effectiveness of a reflective vs. a directive coaching approach on teachers' perceived utility of the automated feedback?
3. When integrated with automated feedback, what is the relative effectiveness of a reflective vs. a directive coaching approach on students' perceived cognitive engagement in math lessons and sense of classroom belonging?

Secondary:

1. When integrated with automated feedback, what is the relative effectiveness of a reflective vs. a directive coaching approach on  student learning outcomes?

**Data Collection Plan and Approach:**

Teachers will audio and video record six mathematics lessons each via high quality Swivl devices to capture snapshots of their teaching throughout a school year. The mathematics teacher - and any support staff members in the classroom - will wear a microphone on a lanyard to clearly capture teacher speech. There will be an additional four microphones scattered around the classroom to capture student speech. Based on initial piloting, we expect to capture nearly 100% of teacher utterances and 90% of student talk using automatic speech recognition (ASR) software; we then use these transcripts to generate teacher feedback. Any incompleteness in this data will then be corrected by human transcriptionists for the purpose of creating our outcome measures, described below. We will also collect administrative data on teachers and students directly from district partners. We will then combine these different elements together to form our analytic dataset.

Measures:

For this experiment, we will focus teachers' attention on three related measures, all with the potential to create more equitable classrooms. Our team has created these NLP-based measures, reported their psychometric properties, and used such measures as automated feedback in prior work (Demszky et al., 2021; Demszky et al., 2023; Demszky & Liu, 2023; Demszky & Hill, 2023; Demszky et al., 2024 ). These measures are:

- *Student reasoning*. This measure identifies student reasoning via the grammatical constructions students frequently use in moments of reasoning (e.g., if, because, while, I think, the way that I know, probably). We focus on student reasoning in this study because it correlates to student outcomes and is a critical component of both current standards-based reforms as well as culturally responsive mathematics teaching, with this measure dovetailing with the latter's emphasis on academic rigor (Ladson-Billings, 1995). The technical details and validity of this measure can be found in Demszky & Hill (2023).

- *Teacher uptake*. This measure identifies instances in which teachers build on the contribution of their students by, for example, acknowledging, repeating, or reformulating what they have said (e.g., Student: "I added 30 to 70." Teacher: "Where did the 70 come from?"). We focus on this measure because it correlates with student outcomes and is a natural follow-on to student reasoning. For example, after increasing the amount of student reasoning, coaches in our pilot often work with teachers on how to take up and build on that reasoning. Teachers' uptake of student ideas promotes dialogic instruction by amplifying student voices and giving them agency in the learning process, unlike monologic instruction where teachers lecture at students (Bakhtin, 1981; Wells, 1999; Nystrand et al., 1997). Our team has constructed an NLP-based measure on teacher uptake using an unsupervised learning approach. We have published a paper that documents the technical details of our NLP approach and the validity of this measure (Demszky et al., 2021).
- *Focusing questions.* Questioning happens all throughout a math lesson. The metric on focusing questions attends specifically to teacher questioning that is meant to open up space for student talk and reasoning. Focusing questions probe students to voice their ideas, reflect on their own or other students' thinking, and to deepen their understanding of the mathematics. When teachers ask focusing questions, they treat students' contributions as valuable ideas for further exploration and sense making. This contrasts with asking questions that direct students towards a desired solution path without much attention to the students' method and reasoning. The technical details and validity of this measure can be found in Alic, et al., 2022.

In addition, we will collect four types of outcomes:

- Teacher surveys after each lesson about the utility of feedback as well as perceptions about changes in student reasoning and students' formation of positive mathematical identities. See appendix A for the specific survey items.
- Automated measures of student reasoning, teacher uptake, and student idea attribution.
- Brief student surveys after each lesson that capture a) students' perception of cognitive engagement in each lesson (a dependent variable that corresponds to student reasoning); and b) students' sense of classroom belonging (a dependent variable that should respond to uptake and student idea attribution). See appendix B for the specific survey items.
- Student test score outcomes from state standardized tests. We will also collect student prior year's test scores as baseline controls.

Note that because the first three types of outcomes are measured at the lesson level, we will have longitudinal data. We will discuss how we leverage this feature of the data when we describe our analytic approach.

**Analytic Approach:**

*Balance Check:* We will start our analysis by running balance tests to check whether our randomization is implemented successfully. We will use teacher characteristics, student characteristics (aggregated to the teacher level), and baseline teaching practices (e.g., teacher uptake) from the first recordings as outcomes in this analysis. These baseline characteristics will also serve as control variables in our analysis of the relative effectiveness of the two coaching approaches.

As described above, because our randomization of teachers will be conducted at the coach level for each of the two cohorts of teachers, we will control for a coach-by-cohort fixed effect in these regressions and cluster the standard errors at the coach level. Besides checking balance for each individual covariate, we will also run a joint test to examine whether there are any systematic differences between the two conditions.

*Regression Models:* For all our primary research questions, we will have repeated measures at the lesson level. For our secondary question, we will only have student test scores at the end of a school year. Thus, we will conduct two types of analyses. First, for RQs 1-3, we will aggregate the lesson-level measures to the teacher level. We will then estimate the overall differences of the two types of coaching for all four research questions. Specifically, we will estimate a model in the following form for teacher-level outcomes (i.e., teacher perceptions and teaching practices):

$$y_i = \beta_0 + \beta_1 T_i + \theta_{ct} + \beta_2 X_i + \varepsilon_i \qquad (1)$$

where $y_i$ indicates teacher ’s outcome such as their teaching practices and their responses to the teacher survey; $T_i$ is a binary indicator for the treatment condition; $\theta_i$ indicates the coach-by-cohort fixed effects; $X_i$ is the rich covariates we described above, including teacher characteristics, student demographics and characteristics (aggregated to the teacher level), and baseline teaching practices (e.g., teacher uptake) from the first recordings which are collected before any coaching interventions; and $\varepsilon_i$ is the error term. $\beta_1$ captures the differential effects on teacher-level outcomes comparing the two coaching approaches. Similar to our balance tests, we will cluster the standard errors at the coach level. Given we have many outcomes to examine in RQs 1-3, we will create composite scores for the student and teacher surveys. To avoid false positives, we will also control the false discovery rate by using Benjamini-Hochberg procedure. When we expand the analysis to student outcomes, we will conduct the regression at the student level to increase our statistical power.

Second, for our primary research questions 1-3, we will leverage the rich longitudinal data we will collect and model the growth of teaching practices to capture the dynamics of how instruction evolves differently comparing the two coaching approaches. We will use a growth curve model that evaluates how teachers’ practices evolve over time, an approach widely used in settings like ours when there are repeated measures for the same individuals over time (Singer & Willett, 2003). We will start with linear growth trajectories first and also examine the possibilities of nonlinearity.

**Power Analysis:**

Our planned sample size is 180 teachers and 4,500 students. Assuming each coach works with 4 teachers at a time in the two cohorts, we will work with approximately 22-23 coaches in total. A power analysis indicates we will achieve 80% statistical power (two-tailed alpha = 0.05) for a minimal detectable effect size (MDES) of 0.42 for the teacher-level outcomes (i.e., instructional practices) and 0.20 for the student-level outcomes (i.e., test scores). Here we assume the proportion of variance in teacher and students outcomes explained by block and relevant covariates is 50%. The MDES for teacher-level outcomes is consistent with prior studies that examine teachers’ instructional outcomes (e.g., Garet et al., 2016; Kraft & Hill, 2020).

**References:**

Alic, S., Demszky, D., Mancenido, Z., Liu, J., Hill, H., & Jurafsky, D. (2022). Computationally identifying funneling and focusing questions in classroom discourse. arXiv preprint arXiv:2208.04715.

Demszky, D., Liu, J., Mancenido, Z., Cohen, J., Hill, H., Jurafsky, D., & Hashimoto, T. (2021). Measuring conversational uptake: A case study on student-teacher interactions. arXiv preprint arXiv:2106.03873.

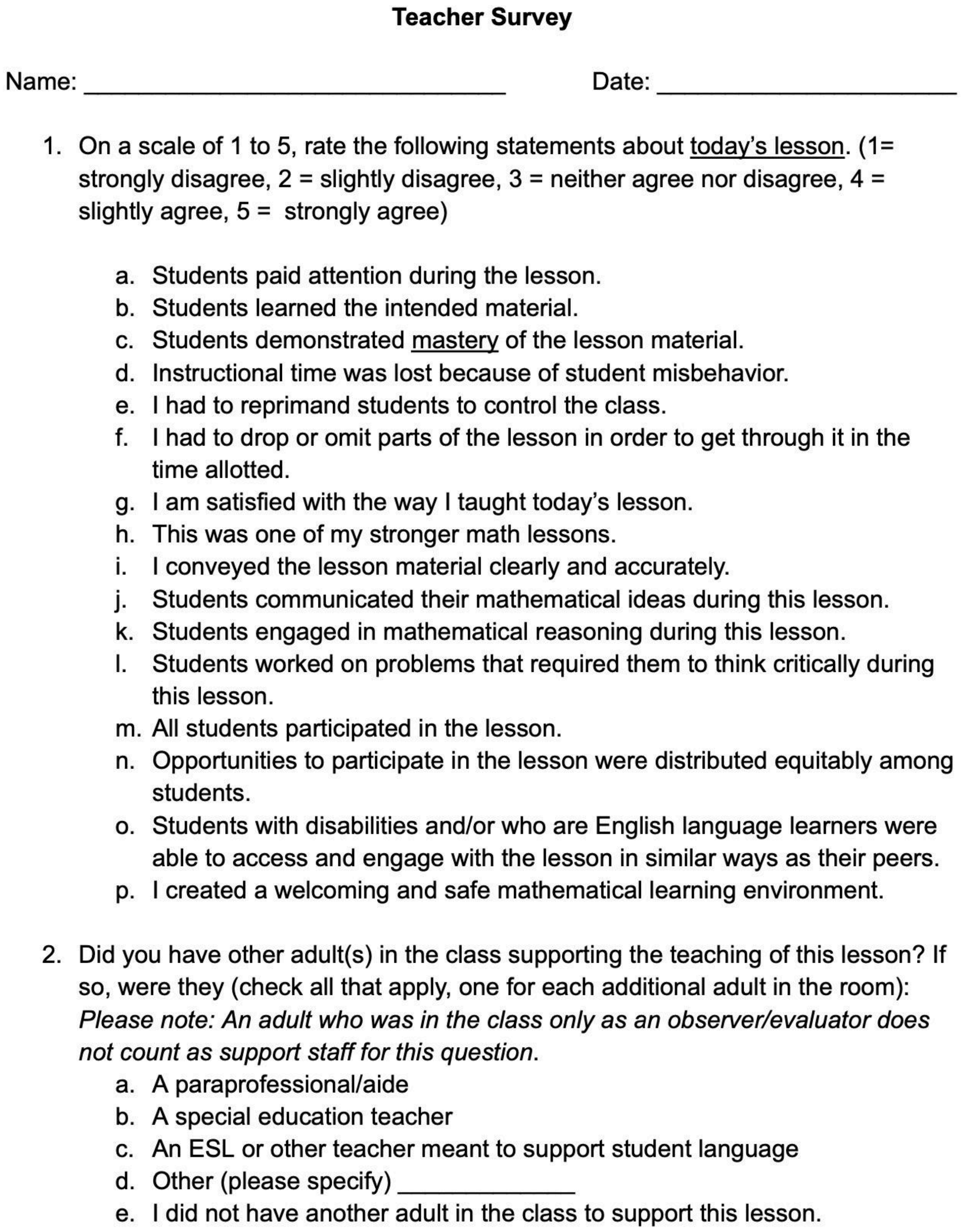
Demszky, D., Liu, J., Hill, H. C., Sanghi, S., & Chung, A. (2024). Automated Feedback Improves Teachers' Questioning Quality in Brick-and-Mortar Classrooms: Opportunities for Further Enhancement. Computers & Education, 105183.

Demszky, D., & Liu, J. (2023, July). M-powering teachers: Natural language processing powered feedback improves 1: 1 instruction and student outcomes. In Proceedings of the Tenth ACM Conference on Learning@ Scale (pp. 59-69).

Demszky, D., Liu, J., Hill, H. C., Jurafsky, D., & Piech, C. (2021). Can Automated Feedback Improve Teachers' Uptake of Student Ideas? Evidence from a Randomized Controlled Trial in a Large-Scale Online Course. EdWorkingPaper No. 21-483. Annenberg Institute for School Reform at Brown University.

Demszky, D., & Hill, H. (2023). The NCTE transcripts: A dataset of elementary math classroom transcripts. 18th Workshop on Innovative Use of NLP for Building Educational Applications.

Appendix A: Teacher Survey



## Teacher Survey

Name: ______________________________ Date: _____________________

1. On a scale of 1 to 5, rate the following statements about today's lesson. (1= strongly disagree, 2 = slightly disagree, 3 = neither agree nor disagree, 4 = slightly agree, 5 = strongly agree)

    a. Students paid attention during the lesson.
    b. Students learned the intended material.
    c. Students demonstrated mastery of the lesson material.
    d. Instructional time was lost because of student misbehavior.
    e. I had to reprimand students to control the class.
    f. I had to drop or omit parts of the lesson in order to get through it in the time allotted.
    g. I am satisfied with the way I taught today's lesson.
    h. This was one of my stronger math lessons.
    i. I conveyed the lesson material clearly and accurately.
    j. Students communicated their mathematical ideas during this lesson.
    k. Students engaged in mathematical reasoning during this lesson.
    l. Students worked on problems that required them to think critically during this lesson.
    m. All students participated in the lesson.
    n. Opportunities to participate in the lesson were distributed equitably among students.
    o. Students with disabilities and/or who are English language learners were able to access and engage with the lesson in similar ways as their peers.
    p. I created a welcoming and safe mathematical learning environment.

2. Did you have other adult(s) in the class supporting the teaching of this lesson? If so, were they (check all that apply, one for each additional adult in the room): *Please note: An adult who was in the class only as an observer/evaluator does not count as support staff for this question*.
    a. A paraprofessional/aide
    b. A special education teacher
    c. An ESL or other teacher meant to support student language
    d. Other (please specify) _____________
    e. I did not have another adult in the class to support this lesson.

3. If you answered (a-d) for the previous question, what did the other adult(s) do during the lesson? Check all that apply.
   __ Supported individual student(s) academically while I taught the lesson
   __ Supported individual student(s) behaviorally/socially/emotionally while I taught the lesson
   __ Worked with a small group of students
   __ Co-taught the lesson to the whole class
   __ Provided special education and/or ELL services
   __ Worked with students in an alternate location (e.g., the hallway or another room)
   __ Other (please specify) ___________________________________
   __ N/A

4. Please upload any materials used during today's lesson (picture or PDF of the lesson, slide deck, worksheets, reference sheets, assessments, etc.)

5. How many previous years have you used these lesson materials? Please count prior years even if you made adaptations.
   a. This was my first time using these lesson materials.
   b. I've used these lesson materials once before.
   c. I've used these lesson materials 2 times before.
   d. I've used these lesson materials 3 or more times before.

6. Anything else that you would like to add about today's lesson?

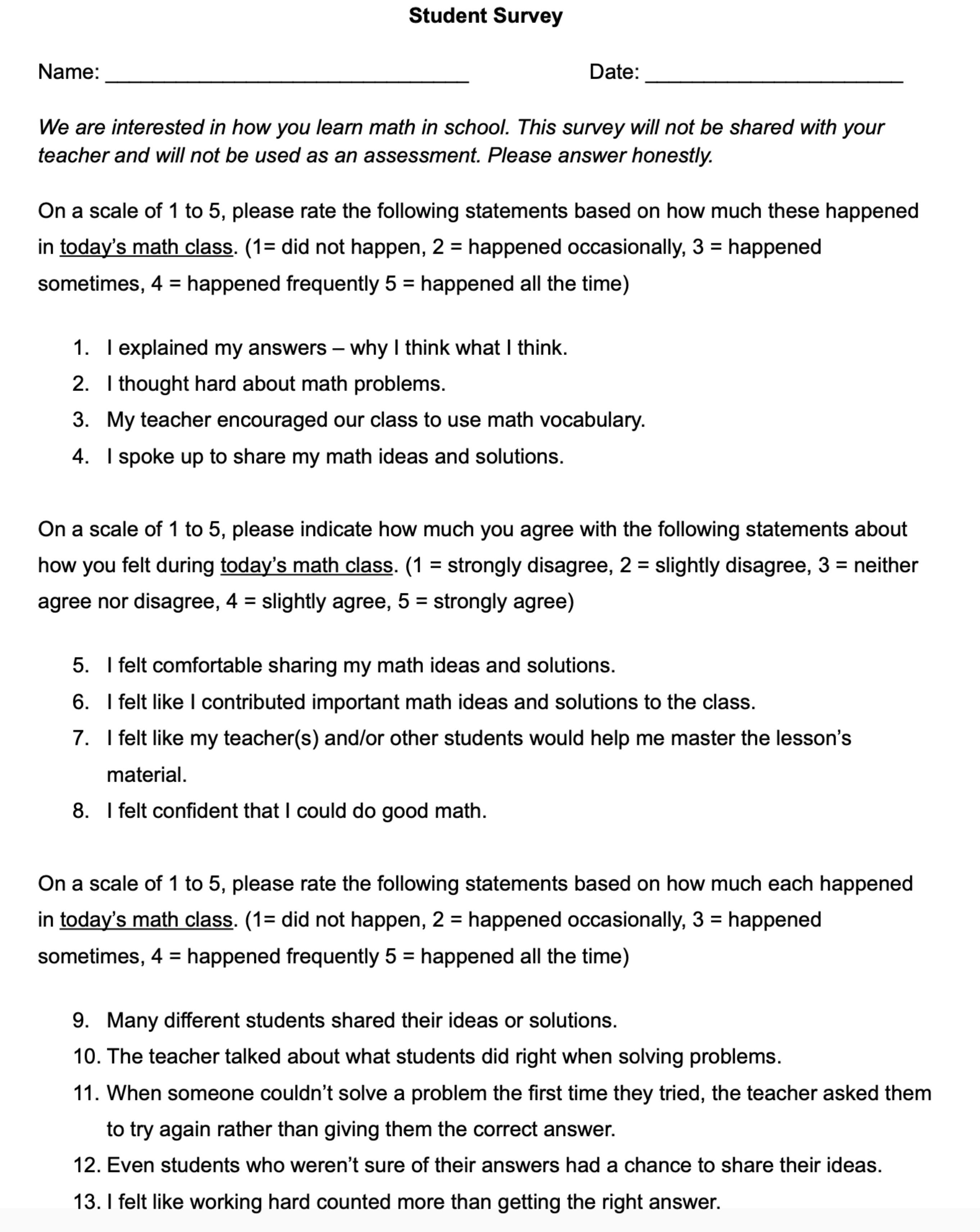
---

*The previous questions will be included each time the teacher takes the survey. The following questions will only be added to the final survey, after teachers have completed all coaching conversations:*

7. On a scale of 1 to 5, answer the following questions based on the coaching you received as part of this study. (1= not much at all, 2 = a little bit, 3 = some, 4 = quite a bit, 5 = a tremendous amount)

   a. Overall, how much do you think your math instruction has improved since your initial coaching conversation?
   b. How much did you change your mathematics teaching based on new ideas/techniques you learned from your coach?

## Appendix B: Student Survey



### Student Survey

Name: ______________________________ Date: _____________________

*We are interested in how you learn math in school. This survey will not be shared with your teacher and will not be used as an assessment. Please answer honestly.*

On a scale of 1 to 5, please rate the following statements based on how much these happened in today's math class. (1= did not happen, 2 = happened occasionally, 3 = happened sometimes, 4 = happened frequently 5 = happened all the time)

1. I explained my answers – why I think what I think.
2. I thought hard about math problems.
3. My teacher encouraged our class to use math vocabulary.
4. I spoke up to share my math ideas and solutions.

On a scale of 1 to 5, please indicate how much you agree with the following statements about how you felt during today's math class. (1 = strongly disagree, 2 = slightly disagree, 3 = neither agree nor disagree, 4 = slightly agree, 5 = strongly agree)

5. I felt comfortable sharing my math ideas and solutions.
6. I felt like I contributed important math ideas and solutions to the class.
7. I felt like my teacher(s) and/or other students would help me master the lesson's material.
8. I felt confident that I could do good math.

On a scale of 1 to 5, please rate the following statements based on how much each happened in today's math class. (1= did not happen, 2 = happened occasionally, 3 = happened sometimes, 4 = happened frequently 5 = happened all the time)

9. Many different students shared their ideas or solutions.
10. The teacher talked about what students did right when solving problems.
11. When someone couldn't solve a problem the first time they tried, the teacher asked them to try again rather than giving them the correct answer.
12. Even students who weren't sure of their answers had a chance to share their ideas.
13. I felt like working hard counted more than getting the right answer.