Measuring Equitable Science Instruction at Scale

Andrew E. Krumm, University of Michigan, aekrumm@umich.edu
William R. Penuel, University of Colorado Boulder, william.penuel@colorado.edu
Carol S. Pazera, University of Texas Austin, carol.pazera@austin.utexas.edu
Carolyn J. Landel, University of Texas Austin, carolynlandel@austin.utexas.edu

Abstract: Using data from the first year of a multi-state curriculum development and implementation initiative, this paper details a quantitative approach to monitoring patterns of equity and inequity in classroom learning opportunities. At a classroom level of analysis, we used instructional logs and a multilevel latent variable model to explore variation in the science practices that teachers enacted and to quantify the relationship between the enactment of science practices and classroom demographics. We used an electronic exit ticket and two multilevel logistic regression models to estimate differences in students' contributions to knowledge-building activities in the classroom controlling for student-level demographics. These analyses demonstrate that the instruction students experience can be measured at multiple levels of analysis and that these analyses can reveal differences that are associated with the race, gender, and home language of students.

Introduction

Promoting equity is a priority for science education. In an equitable classroom, students' experiences should not be predictable simply by knowing their race, self-identified gender, language, or disability status. Providing meaningful experiences to students from diverse backgrounds not only can prepare students for civic engagement within a more just, democratic society (Morales-Doyle, 2017), it can also help to diversify and thus enrich science (Medin & Bang, 2014). Yet, educational systems and instructional programs working to achieve equity face a number of challenges, and *measurement* is central among these challenges in promoting equity. To best understand and monitor progress toward equity goals, researchers, practitioners, and policymakers need measures that are valid, reliable, and practical for administering at the scale of schools, districts, and states (Herman, Klein, & Abedi, 2000; Mislevy, 2008). Proposed measures, such as access to advanced course taking (e.g., NASEM, 2019), are useful for understanding variation in educational opportunities, but these proposals provide little insight into what occurs inside of classrooms, where inequalities can easily be reproduced within classroom interactions (e.g., Solís et al., 2009). In fact, few—if any—measurement approaches are designed to assess instructional equity. Within the context of science classrooms, instructional equity can be thought of as opportunities for students to participate in and contribute to learning activities organized around science and engineering content in ways that connect to student interest and that advance students' local community priorities (Tan & Calabrese-Barton, 2012). Of particular importance is attending to how instruction can support participation and contributions of students that have been and continue to be underrepresented in science and engineering because of their race, gender, language, or ascribed abilities (Penuel & Watkins, 2019; Philip & Azevedo, 2017).

Recent efforts to establish a more robust form of *equity analytics* (Reinholz & Shah, 2018) are moving in promising directions. Reinholz and Shah intend equity analytics to be a quantitative complement to in-depth, qualitative, nuanced examinations of patterns of equity and inequity inside classrooms. Their study shows how observational data can reveal ways that inequities of opportunity to participate in rich classroom discourses can emerge, even in equity-minded teachers' classrooms. Similarly, Munter and Haines (2018) illustrate the potential of equity analytics to investigate the relation of instruction to students' racialized experiences of classrooms, showing ways that the cognitive demand of tasks in mathematics are differentially perceived by White and racially minoritized students. While these approaches are promising, there is a strong need among researchers, practitioners, and policymakers for valid and reliable measures of instructional opportunities that can be collected, disaggregated by student demographics, and yield meaningful inferences at scale.

In this paper, we explore the use of two data collection instruments—instructional logs completed by teachers and electronic exit tickets completed by students—to quantify students' experiences of science instruction. The logs and exit tickets were administered multiple times throughout the academic year during a 10-state field test of new middle school instructional materials as part of OpenSciEd. During the 2018-19 school year, 6 newly developed and refined science units were field-tested (1 in the fall and 1 in the spring in grades 6, 7, and 8). OpenSciEd is comprised of a consortium of developers (i.e., both instructional materials and professional learning experiences) and state science leads that is organized around providing free, open, and high-quality resources and implementation supports. The field test involved collecting and analyzing data for OpenSciEd developers and state leads in order to improve instructional materials and professional learning experiences over time; therefore, data were collected with the express purpose of identifying potential concerns as well as bright spots during the field test. As part of this improvement-focused work, most analyses were descriptive in nature and pertained to the units that were, at the time, being implemented by teachers. In this paper, we describe in-depth statistical analyses using the entire first years' worth of data to understand variation in science instruction and the degree to which that variation was explained by three aspects of students' identities, namely, their self-identified race, gender, and home language. We use previous research on equity-focused analytics and measuring instructional opportunities to connect our analytical work to implications for the learning sciences.

Theoretical and methodological considerations

A central feature of making classrooms more equitable is attending to various opportunities to learn (OTL), which refers to the quality of curricular content, resources, and teaching made available to students (Guiton & Oakes, 1995; Herman, Klein, & Abedi, 2000). Indicators of OTL make it possible to evaluate the effectiveness of programs and support valid interpretation of assessment data (NRC, 2014). OTL is based on the idea that opportunities provided through instruction are the most proximal causes of student learning in formal learning environments (Cohen, Raudenbush, & Ball, 2003). OTL measures have the potential to support equity goals when they address the degree to which students encounter inclusive instructional approaches and whether such opportunities are distributed equitably. Recent calls for measures of OTL have focused on both discipline-specific measures (NRC, 2014; Wijaya, et al., 2015) and measures that can attend to opportunities of specific groups of students, including English learners and students with disabilities (NASEM, 2019).

Recent research in the learning sciences has focused on the use of structured observations to monitor equity of OTL in classrooms. One such measure, noted previously, is based in Reinholz and Shah's (2018) study of classroom discourse in a single classroom over a 30-hour period using a protocol called Equity QUantified In Participation (EQUIP). Other measures have focused on using observation protocols to document ways that instructors are incorporating culture-based pedagogies into instruction (e.g., Jensen, Grajeda, & Haertel, 2018). Still other approaches have combined observational data with social network analysis to help document how access to teaching expertise is distributed across school districts (Sun, et al., 2014).

Many of these approaches, while useful for research purposes, pose challenges in terms of the speed and scale at which data can be collected, analyzed, and communicated to those who need data for creating more equitable classroom environments. Observational measures, for example, are expensive to collect and demand many resources to score, and reporting quantitative differences by background characteristics can inadvertently treat characteristics of students as a "social address" rather than as a dynamic social construct (Lee, 2002).

A practical approach to measuring equity in opportunities to learn at scale

The importation of improvement science into the education reform landscape has drawn attention to the use of *practical* measures of instruction, that is, measures that can be used easily *in* practice, *about* practice, and *by* educators (Bryk, Gomez, Grunow, & LeMahieu, 2015). Some of these measures build directly on past work to measure OTL, particularly through instructional logs. Instructional logs are short, repeated-measure surveys that teachers complete at the conclusion of a day's lesson throughout the academic year; logs have been shown to be reliable and scalable tools for measuring OTL (Rowan, Camburn, & Correnti, 2004). Other researchers have employed brief surveys targeting students' experience of instruction, which can serve as "exit tickets" collected at the conclusion of instruction (e.g., Ahn, Campos, Hayes, & DiGiacomo, 2019) or as pre-post measures used to evaluate interventions (e.g., Kosovich, Hulleman, Barron, & Getty, 2015).

One example of a strategy for using student surveys to measure equity is the Student Electronic Exit Ticket (SEET) used by the InquiryHub research-practice partnership in Denver. Teachers administer the Google Forms-based exit ticket every 2-3 weeks throughout the school year to students. Each exit ticket asks students to report on the perceived coherence of the lesson, the personal and community relevance of the lesson, as well as the degree to which they contributed in class and whether those contributions were recognized by others. Data on self-identified race, gender, and home language are also collected. The data are intended to help the partnership monitor equity of opportunity at the district level, by analyzing variation in student experience within and across classrooms by background characteristics with respect to coherence, relevance, and contribution (Penuel & Watkins, 2019). Research on the SEET indicates student experience varies markedly from classroom to classroom (Penuel, Van Horne, Severance, Quigley, & Sumner, 2016) and that visualization of SEET data can help teachers attend to inequities within their classrooms (Raza, Penuel, Jacobs, & Sumner, in press).

Building on equity analytics, OTL measurement, and improvement-focused examples, OpenSciEd set out to understand, monitor, and support the improvement of equitable science instruction. As part of this emphasis on equity, data were collected from samples of teachers across units using a variety of approaches that included instructional logs and exit tickets as a practical and scalable approach. The analyses described below extend upon data products that were regularly presented to OpenSciEd developers and state leaders. Our modeling approach seeks to complicate overly simplistic narratives about how students' experience in the classroom varies as a function of their "social address" by highlighting within-group variation and investigating variability linked to classroom composition (i.e., more heterogeneous versus less heterogeneous classrooms).

Methods

In this section, we describe our instruments, sample and procedures, and analytical approach for understanding teacher- and student-level variation in students' experience of instruction and the degree to which variation in instruction is attributable to student characteristics. Analyses were organized around the following research questions: (1) How variable is teachers' self-reported enactment of instruction, and does it vary by classroom composition? (2) How variable are students' contributions to knowledge-building activities in the classroom, and do students' contributions vary based on their background?

Instruments

Teacher logs

Teacher logs included two sets of questions. One set related to teachers' instructional moves for supporting students' experience of the coherence of lessons with the overall focus of the unit, that is, the unit's anchoring phenomenon. An anchoring phenomenon serves as part of the backbone for each unit; students build explanatory models of the anchoring phenomenon over the course of the unit, using core ideas, crosscutting concepts, and science practices (Penuel & Reiser, 2018). Teachers answer "yes" or "no" to the following questions related to coherence-building, student-facing practices on that day's lesson: *At the beginning of class, students discussed what we figured out during the previous lesson*; *Students discussed a connection between the focus of today's lesson and the anchoring phenomenon*; and *Students discussed what they figured out at the end of the lesson.* A second set of three items relate to supports for student sensemaking (Odden & Russ, 2019) and use the same "yes-no" answer format: *Students revised claims they had made earlier based on evidence or information gathered during the day's lesson*; *Students generated gaps or inconsistencies in their explanatory models of a phenomenon*; and *Students generated claims related to the day's investigation on the basis of what they have investigated together.*

Student exit tickets

For the purposes of the above research questions, we focus on a pair of items from an exit ticket that was administered by teachers to students following a lesson. Students completed the exit ticket electronically. Each exit ticket was anonymous, which meant that students self-identified key demographic information in addition to answering questions related to coherence, relevance, and classroom contributions. We analyzed two items from the exit ticket related to contribution, which refers to students' perceptions that they are contributing ideas to knowledge building in the classroom. Contributing toward knowledge building is an important aspect of a student's capacity to shape and evaluate knowledge-building practices in classrooms. The two items students answered were: *Did you share any ideas out loud today to the whole class, a small group or a partner?* and *If you answered yes to the last question, did any of your ideas influence the class or help others?* Students responded "yes" or "no" to these questions. We analyzed 8,178 logs across 75 teachers from 5 OpenSciEd units.

Sample and procedures

A unique feature of OpenSciEd was an intentional focus on equity during the first year of the development initiative. Given the early nature of the work, an explicit choice was made to reduce data collection burden on teachers and to have different samples of teachers collect logs and exit tickets. Each sample represented approximately 25% of all participating teachers per unit. Across all OpenSciEd teachers (N = 259), approximately 90% self-identified as female, and the mean number of years taught was 12.8 with the median teacher having taught 12 years. Approximately 82% of teachers self-identified as White (N = 213), 7.3% as Native Hawaiian or Pacific Islander (N = 19), 3.5% as multiracial (N = 9), and 3.1% as African American or Black (N = 8), with 10 teachers not reporting a race or ethnicity. Teacher background characteristics were similar across data collection groups.

Teachers in both data collection groups were assigned four specific lessons from which to gather data, which helped in organizing the data collection process because teachers could implement a fall or spring unit anytime during the fall or spring. Emails were regularly sent to teachers reminding them of their assigned data collection group, relevant lessons, and links to data collection instruments. A website for participating teachers was used to organize the data collection process. Other data that were collected included teachers' self-reported classroom demographics for a focal class from which they administered logs or exit tickets.

For the analyses described below, we report on data collected from five units: three in the fall and two in the spring. A possible third spring was not included because fewer teachers implemented it overall as compared to the other five units. Some teachers participated in more data collection activities than they were assigned for a given unit; only teachers who were assigned to a group were included for analysis. We analyzed 237 logs for 77 teachers across 5 OpenSciEd units. On average, teachers submitted approximately 3 logs and 41 teachers submitted all 4 requested logs. For exit tickets, we analyzed 8,178 individual exit tickets for 75 teachers, and on average, teachers completed 3 exit ticket opportunities with students, and 35 teachers completed all 4 opportunities.

Analytical approach
To understand the variability in teachers' self-reported enactment of instruction and the degree to which instruction varied by classroom composition (RQ #1), we engaged in a two-step process. First, we fit a multilevel latent variable model to generate a teacher-level score (De Boek, et al., 2011). The teacher-level score is similar to the one outlined by Rowan, Camburn, and Correnti (2004): it quantified the log-odds of whether a teacher enacted a practice within a particular lesson. We nested individual responses within lessons and within teachers, which allowed us to quantify how the enactment of practices varied by teacher and lesson. For the second step in our process, we then compared teacher-level scores with data reported by teachers on their focal classroom's composition. We operationalized classroom composition data as a concentration variable based on the maximum percent represented by a single race or ethnicity—the higher the maximum value, the more homogenous the class. We tested the influence of the concentration variable using a separate linear regression model. In constructing this two-step analysis, positive evidence for equitable instruction across the sample of teachers would entail no significant relationship between homogeneity and the likelihood of enacting one or more science practices. Homogeneity was our primary operationalization of classroom-level demographics in an effort to avoid using White students as the reference group and unwittingly reinforcing problematic approaches to improvement that focus on "bringing up" minoritized students to a particular standard (Gutiérrez, 2008).

For exit tickets, we examined student-level identifications of race, gender, and home language. Student-level identifications afforded the opportunity to examine whether students with the same background characteristics experienced instruction differently. We estimated two general multilevel models for the two contribution questions described previously. Student responses were nested within lessons and within teachers. We used the variance components from the models to quantify the variation in students' experience of instruction across teachers and lessons. Student-level fixed effects included dummy variables for students' self-reported ethnicity: African American or Black, American Indian or Native American, Asian or Asian American, Latinx or Hispanic, multiple, Native Hawaiian or Pacific Islander, and White. Checks were done to compare aggregate student reports and teacher-provided classroom-level values, and no dramatic differences were detected. We also included a classroom-level homogeneity variable and two-way interactions between the homogeneity variable and student-level ethnicity variables . Other fixed effects included gender and language spoken at home: English, multiple, Spanish, and other. Given the potential for collinearity among demographic variables, we examined correlations among independent variables, and the highest correlation was between Latinx students and Spanish as the primary language spoken at home ($r = .64$, $p < .001$), which meant that multicollinearity was not a concern.

Results

How variable is teachers' self-reported enactment of instruction, and does it vary by classroom composition?
Table 2 presents the results from the first step in our two-step analytical process. Step one involved generating a teacher-level score from the multiple practices enacted within a lesson. Four of six practices were enacted with high regularity (i.e., teachers enacted these practices in over 80% of lessons). However, when analyzed as combinations of practices enacted across lessons, we observed variability in instruction, whereby relatively few teachers enacted each practice across multiple lessons. Lastly, we did not identify a significant relationship, either positive or negative, between instruction and equity at the teacher-level of analysis.

Table 2 presents both the percent of lessons where a teacher self-reported enacting a coherence-building or sensemaking practice, the model-based probability that a teacher enacted a practice, and the estimated log-odds, conditional on teacher and lesson, that a teacher enacted a practice. Across the multiple ways of understanding how likely a practice was enacted, results illustrate that coherence-building practices had a high likelihood of being enacted on a lesson-by-lesson, teacher-by-teacher basis. Two sensemaking practices, on the other hand, were less likely to be enacted: *Students revised claims they had made earlier based on evidence or information gathered during the day's lesson* and *Students generated gaps or inconsistencies in their explanatory models of a phenomenon*.

The random effects from the model point to potential sources of variation among different practices. More than twice as much variation exists at the teacher level of analysis than at the lesson level. This means that individual teachers differed from one another in terms of the practices they enacted from lesson to lesson. The strength of that difference, however, can be thought of as low. Using the estimated variance components, the correlation among practices for the same teacher across *different* lessons was approximately .23, which signals that specific combinations of practices were not consistently enacted.

Table 1. Instructional log multilevel latent construct modeling results

	% Lessons	Estim. Prob.	*B*	Std. Error
Coherence-building practice				
Figured out during previous lesson	90.5%	.94	2.778 ***	.285
Connection to anchoring phenomenon	80.7%	.86	1.796 ***	.233
Figured out at end of lesson	82.3%	.87	1.928 ***	.238
Sensemaking practice				
Revised claims during lesson	72.0%	.77	1.189 ***	.214
Generated gaps or inconsistencies in models	60.2%	.62	.497 *	.202
Generated claims on what was investigated	84.4%	.89	2.108 ***	.246
Random effects			Variance	
Lesson			.468	
Teacher			1.118	

Note: *** = $p < .001$; ** = $p < .01$; * = $p < .05$

A benefit of the latent construct modeling approach that was used is that a teacher-level random effect can be extracted and used to quantify the likelihood that a teacher enacted a specific practice. Using this teacher-level score, we examined the relationship between teachers' likelihood of enacting one or more practices with teachers' overall classroom composition operationalized in terms of demographic homogeneity. A separate regression model—step two in the analytical approach used for this research question—comparing classroom homogeneity with the teacher-level score produced a near-zero relationship that was not statistically significant (B = .004, p = .41). Along with regressing classroom homogeneity on the teacher-level score, we also experimented with various operationalizations of classroom-level demographics (e.g., percent of students from non-dominant groups), and similarly, did not find any relationship between teachers' enactment of instruction and classroom composition. We also conducted follow up analyses, repeating the same two-step process as above, on the "generating gaps or inconsistencies" practice, alone, as it was the least frequently implemented practice, which made it potentially indicative of a practice that could be mediated by classroom composition. Subsequent analyses revealed a slightly stronger relationship between homogeneity and the teacher-score, but it was weak and not significant at the $p < .05$ level (B = .110 , p = .09).

How variable are students' contributions to knowledge-building activities in the classroom, and do students' contributions vary based on their background?

Exit tickets provide a different window into the experiences of students. Based on descriptive and statistical analyses, we found that race was a significant predictor for both contribution measures, and in particular, for overcoming the hurdle of contributing to class in the first place. There was also an effect of gender for sharing ideas, but not for having ideas heard, with males and gender nonbinary students reporting that they share less often. There were no significant effects of home language on either contributing to class or influencing peers. Unlike for the log sample of classrooms, classroom-level homogeneity was a significant predictor as were two interactions between the homogeneity variable and student-level fixed effects for Asian students and students who self-identified as mixed race.

Table 2 presents descriptive statistics for submitted exit tickets by background characteristics and the two knowledge-building items: contribution and influence. The percent of students who responded "yes," represents the baseline for understanding more complex modeling results presented in Table 3. In Table 2, Asian and White students reported contributing to class with the highest frequency, 69% and 70%, respectively.

Hawaiian and Pacific Islander students and American Indian students self-report contributing to class with the lowest frequency, 51% and 60%, respectively. It is important to recognize that, with the exception of Hawaiian and Pacific Islander students, the range between students from academically marginalized (e.g., African America and Latinx) and non-marginalized (e.g., White and Asian) backgrounds is relatively small at approximately 10%. Yet, the results presented in Table 3 add more precision and integrate more rigorous controls in understanding variability in students' participation and the effects students' background characteristics.

Table 2. Descriptive statistics for students' self-reported background characteristics on exit tickets

	Contribution			Influence		
	Count	Yes %	No %	Count	Yes %	No %
Race / Ethnicity						
Latinx	1,258	61	39	763	73	27
Asian	456	70	30	316	71	29
Multiple	1,265	65	35	816	74	26
African Amer.	723	64	36	463	74	26
Amer. Indian	265	60	40	159	71	29
Hawaiian / P.I.	37	51	49	19	63	37
White ^	4,204	69	31	2,883	74	26
Gender						
Male	3,809	64	36	2,437	73	27
Non-binary	42	52	48	22	64	36
Female ^	4,042	70	30	2,804	75	25
Home Language						
Multiple	780	64	36	498	75	25
Other	418	62	38	255	71	29
Spanish	566	60	40	341	76	24
English ^	6,314	68	32	4,296	74	26

Note: ^ = reference category for models 1 and 2 in Table 2.

Table 3. Student exit ticket models for contributing to knowledge-building activities

	Model 1			Model 2		
	B	Std. Error	*Z*	*B*	Std. Error	*Z*
Intercept	1.030	.082	12.50 ***	1.136	.078	14.56 ***
Race / Ethnicity						
Latinx	-.339	.099	-3.41 ***	-.337	.131	-2.58 *
Asian	.037	.133	.28	-.190	.165	-1.16
Multiple	-.172	.076	-2.52 *	-.034	.099	-.35
African Amer.	-.119	.095	-1.25	-.056	.123	-.46
Amer. Indian	-.376	.148	-2.55 **	.223	.224	1.00
Hawaiian / P.I.	-.702	.350	-2.01 *	-.566	.515	-1.10
Gender						
Male	-.259	.050	-5.21 ***	-.076	.065	-1.81
Non-binary	-.724	.325	-2.23 *	-.523	.464	-1.13
Home Language						
Multiple	-.117	.095	-1.23	.192	.129	1.49
Other	-.240	.124	-1.94	-.059	.164	-.36
Spanish	-.125	.120	-1.04	.300	.169	1.78
Teacher-Level						
Homogeneity	.210	.079	2.66 **	.022	.072	.30
Asian:HG	-.289	.114	-2.54 *			
Multiple:HG	-.154	.076	-2.02 *			
Random effects	Variance			Variance		
Lesson	.094			.060		
Teacher	.279			.146		

Note: Only significant two-way interactions are reported. *** = $p < .001$; ** = $p < .01$; * = $p < .05$.

Table 3 illustrates how students who identify as Latinx, Multiple, American Indian, and Hawaiian or Pacific Islander are less likely to contribute than White students. Importantly, for Model 2, we observe less of an influence of demographics on whether students perceive their contributions as influencing the class. The only main effect that was significant was for students who identify as Latinx. While there were significant differences at the student-level of analysis, classroom homogeneity was also a significant predictor in Model 1. The direction of the effect indicates that the more homogenous the class, the more likely students are to contribute. However, there were negative interactions for students who self-identified as Asian or Multiple, meaning that the more homogenous the classrooms, the less likely these groups of students were to contribute. In further exploring classroom-level factors, we compared classroom proportions for groups of students with predicted probabilities from Model 1 for the same group of students. These analyses revealed that the probability of students contributing increases as the proportion of students from the same self-identified background increases.

In seeking to better understand teachers' influence on students' contributions, we examined teacher-level random effects from Model 1. Sixteen teachers (~24%) were significantly different from the median teacher, in both positive and negative directions, with the majority of differences occurring on the negative side of the distribution. This means that these teachers had students who were more likely to contribute (or not) to distinctly different degrees than the typical teacher. For five of the highest scoring teachers (i.e., those who engendered more student contributions in their class), three had racially diverse classrooms. In a similar analysis conducted for Model 2, three teachers stood out in terms of having students report that their contributions influenced the class. These three teachers taught in classrooms with a majority of Latinx students. Therefore, even at the student-level of analysis that exit tickets afforded, there were noticeable classroom composition effects on students' likelihood for contributing to classroom knowledge-building activities.

Discussion and conclusion

This paper used two measurement and modeling approaches to understand students' opportunities to learn as well as their experience of instruction during the first year of OpenSciEd. Using data at two levels of analysis, we identified that classroom-level demographics were not significantly related to teachers' instructional practices; classroom composition did matter, however, at a student-level of analysis. A potential takeaway from this overall finding is that efforts to understand students' opportunities to learn and their experience of instruction using an equity lens should, when possible, gather data directly from students. Moreover, student-level experiences should be combined with classroom-level demographics because overall classroom composition and classroom dynamics seem to moderate individual experiences. While we were able to identify that student and classroom factors can affect students' experience of instruction, there are limitations to the strength of the claims that can be made. For example, samples of teachers across instructional logs and exit tickets were not overlapping, which was done to reduce burden on teachers in the service of providing rapid, digestible data products throughout the development and implementation work of OpenSciEd.

The analyses presented in this paper underscore that more work needs to be done to close persistent instructional equity gaps in the face of continuing re-segregation of American schools (Frankenberg, Ee, Ayscue, & Orfield, 2019) and cultural as well as political headwinds (Politics of Learning Writing Collective, 2017). In many ways, the analyses of instructional equity presented in this paper may simply reflect longstanding patterns of inequity in the United States.

Equity is a central focus of OpenSciEd, and for an initiative operating across 10 states, this paper demonstrated ways that other design, development, and implementation projects could measure opportunities to learn while foregrounding equity. An intended outcome of this paper is that more curriculum-focused projects gather data on students' experience of instruction so that comparisons in the variation that we observed can be made and further contextualized. Such comparisons could help in identifying the inequities that predominate in schools from those that are improved or made worse by new instructional materials: our analyses identified opportunities to improve (e.g., student-level classroom contribution differences) and potential bright spots (e.g., teachers in heterogenous classrooms who enacted all measured practices or engendered high degrees of student contributions). Descriptive analyses in this paper highlighted the promising work being done in OpenSciEd as well as the importance of prioritizing instructional equity. In future research, we plan to supplement both descriptive and statistical analyses with more targeted qualitative research in our ongoing efforts to measure equitable science instruction at scale.

References

Ahn, J., Campos, F., Hays, M., & DiGiacomo, D. (2019). Designing in context: Reaching beyond usability in learning analytics dashboard design. *Journal of Learning Analytics, 1*(1), 1-10.

De Boeck, P., Bakker, M., Zwister, R., Nivard, M, Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer functions from the lme4 package in R. *Journal of Statistical Software, 39*(12), 1-28.

Guiton, G., & Oakes, J. (1995). Opportunity to learn and conceptions of educational equality. *Educational Evaluation and Policy Analysis, 17*(3), 323-336.

Gutiérrez, R. (2008). A "gap-gazing" fetish in mathematics education? Problematizing research on the achievement gap. *Journal of Research in Mathematics Education, 39*(4), 357-364.

Herman, J. L., Klein, D. C. D., & Abedi, J. (2000). Assessing students' opportunity to learn: Teacher and student perspectives. *Educational Measurement: Issues and Practice, 19*(4), 16-24.

Jensen, B., Grajeda, S., & Haertel, E. (2018). Measuring cultural dimensions of classroom interactions. *Educational Assessment, 23*(4), 250-276.

Kosovich, J. J., Hulleman, C. S., Barron, K. E., & Getty, S. (2015). A practical measure of student motivation: Establishing validity evidence for the expectancy-value-cost scale in middle school. *Journal of Early Adolescence, 35*(5-6), 790-816.

Medin, D., & Bang, M. (2014). *Who's asking? Native science, Western science, and science education*. Cambridge, MA: MIT Press.

Mislevy, R. J. (2008). Issues of structure and issues of scale in assessment from a situative/sociocultural perspective. In P. A. Moss, D. Pullia, E. H. Haertel, J. P. Gee, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 259-294). New York, NY: Cambridge University Press.

Morales-Doyle, D. (2017). Justice-centered science pedagogy: A catalyst for academic achievement and social transformation. *Science Education, 101*, 1034-1060.

National Academies of Sciences Engineering and Medicine. (2019). *Monitoring educational equity*. Washington, DC: National Academies Press.

National Research Council. (2014). *Developing assessments for the Next Generation Science Standards*. Washington, DC: National Academies Press.

Penuel, W. R., Van Horne, K., Severance, S., Quigley, D., & Sumner, T. (2016). Students' responses to curricular activities as indicator of coherence in project-based science. In C.-K. Looi, J. L. Polman, U. Cress, & P. Reimann (Eds.), *Proceedings of the 12th International Conference of the Learning Sciences* (Vol. 2, pp. 855-858). Singapore: International Society of the Learning Sciences.

Penuel, W. R., & Reiser, B. J. (2018). *Designing NGSS-aligned curriculum materials*. Paper prepared for the Committee to Revise America's Lab Report. Washington, DC: National Academies of Science and Medicine.

Penuel, W. R., & Watkins, D. A. (2019). Assessment to promote equity and epistemic justice: A use-case of a research-practice partnership in science education. *Annals of the American Academy of Political and Social Science, 683*, 201-216.

Philip, T. M., & Azevedo, F. S. (2017). Everyday science learning and equity: Mapping the contested terrain. *Science Education, 101*(4), 526-532.

Politics of Learning Writing Collective. (2017). The learning sciences in a new era of U.S. nationalism. *Cognition and Instruction, 35*(2), 91-102.

Raza, A., Penuel, W. R., Jacobs, J., & Sumner, T. (in press). Supporting equity in schools: Using visual learning analytics to understand students' classroom experiences. In M. Schmidt, A. Tawfik, Y. Earnshaw, & I. Jahnke (Eds.), *Learner and user experience research: An introduction for the field of learning design & technology*: Edtechbooks.

Reinholz, D., & Shah, N. (2018). Equity analytics: A methodological approach for quantifying participation patterns in mathematics classroom discourse. *Journal for Research in Mathematics Education, 49*(2), 140-177.

Rowan, B., Camburn, E., & Correnti, R. (2004). Using teacher logs to measure the enacted curriculum: A study of literacy teaching in third-grade classrooms. *Elementary School Journal, 105*(1), 75-102.

Solís, J., Kattan, S., & Badequeno-López, P. (2009). Socializing respect and knowledge in a racially integrated science classroom. *Linguistics and Education, 20*(3), 273-290.

Sun, M., Wilhelm, A. G., Larson, C. J., & Frank, K. A. (2014). Exploring colleagues' professional influences on mathematics teachers' learning. *Teachers College Record, 116*(4), 1-30.

Tan, E., & Calabrese Barton, A. (2012). *Empowering science and mathematics education in urban schools*. Chicago, IL: University of Chicago Press.

Wijaya, A., van den Heuvel-Panhuizen, M., & Doorman, M. (2015). Opportunity-to-learn context-based tasks provided by mathematics textbooks. *Educational Studies in Mathematics, 89*(41-65).