SN Soc Sci (2021) 1:183
https://doi.org/10.1007/s43545-021-00190-x

ORIGINAL PAPER

Capturing who participates and how: the stability of classroom observations using EQUIP

Daniel L. Reinholz[1] · Kevin Pelaez[2] · Niral Shah[3]

Received: 19 November 2020 / Accepted: 14 June 2021

Abstract

EQUIP is a free, customizable observation protocol for tracking patterns of student participation in STEM classrooms (https://www.equip.ninja). EQUIP generates data analytics that are disaggregated by student social markers (e.g., race, gender), which makes it a useful tool for tracking patterns of inequity in student participation. However, prior studies have not yet established how many observations are needed to create a representative picture of instruction. In this study, we use g-theory and simulations with Cramer's V to analyze observations from 20 undergraduate mathematics instructors to determine how many classroom observations are needed, and how this differs by individual codes. We found that Gender could achieve stability in just a few observations, whereas codes such as Instructor Response, Instructor Solicitation Type, and Instructor Solicitation Method required nearly 20 observations. Thus, we recommend that users account for their specific context and needs with EQUIP when determining the ideal number of observations to conduct, using this research as a baseline. We also compare the g-study and simulations approaches, bringing up new methodological questions for the field.

Keywords Classroom observation · Discourse · Equity · Gender · Race · Teacher education

✉ Daniel L. Reinholz
daniel.reinholz@sdsu.edu

[1] Department of Mathematics and Statistics, San Diego State University, San Diego, CA 92182-7720, USA

[2] Mathematics and Science Education, San Diego State University, San Diego, CA, USA

[3] College of Education, University of Washington, Seattle, WA, USA

Published online: 14 July 2021

Introduction

Classroom observations are ubiquitous in STEM education. Observations are seen as one of the primary ways to authentically capture teaching practices, as compared to instructor self-report or a student survey. As a result, classroom observations are used in a variety of ways: for teacher professional development, for research studies, and for the evaluation of teachers and educational policies. As the use of classroom observations has become increasingly popular, so has the proliferation of classroom observation protocols. Myriad observation tools are now freely available.

Although considerable time, energy, and money is put into classroom observations, the accuracy of classroom observations is still a looming question that plagues the field. In many research studies, a teacher may be observed only one or two times to create a picture of their teaching practice. But how accurate is that picture? Research on classroom observations suggests that many more than two observations are needed to accurately capture teaching (Weston et al. 2020). A variety of factors contribute to the accuracy of an observation, including: the teacher's own characteristics, the subject domain, the grade level, the number of raters, and the particular classroom observation tool. For this reason, teachers working in some contexts may require fewer observations than others to achieve an acceptable level of reliability (Hill et al. 2012). Nonetheless, it is clear that number of observations needed is closely linked to the observation tool in question, and for this reason, a careful analysis of any given classroom observation tool is needed to infer properties about that particular tool.

This manuscript focuses on the classroom observation tool EQUIP (Reinholz & Shah, 2018). EQUIP was designed with an explicit focus on equity, aiming to capture the equitable distribution of participation in classroom discussions. EQUIP has been used in a large number of prior studies both to support professional development (e.g., Reinholz et al. 2019, Shah et al., 2020) and research on classroom inequities (e.g., Ernest et al., 2019; Reinholz & Shah, 2018; Reinholz et al., in press). These studies range from K12 to postsecondary settings, and while many are grounded in mathematics and STEM disciplines, they also focus on non-STEM disciplines as well. Given that EQUIP is a tool for both professional development and research, we hypothesize that a different number of observations may be required to accurately capture each of these purposes. In general, more observations are required for research purposes as compared to professional development. This paper addresses the following research questions:

1. How many observations are required to accurately capture teaching practices with EQUIP?
2. How does the number of observations differ by EQUIP codes?

Drawing on an existing dataset from a single large-scale research study in undergraduate mathematics (Reinholz et al. in press), we characterize the stability of classroom observations with EQUIP across 20 teachers in undergraduate

mathematics by using generalizability theory (Brennan 2001), and a simulation, to identify the optimal number of observations needed to accurately capture teaching practices.

Background

Stability of classroom observations

Over the last decade, considerable effort has been put into understanding the stability of classroom observations. There are a number of reasons for this. First, classroom observations are relatively time consuming and expensive, so it is ideal to limit the number of observations needed for any given purpose. Second, the results of observations can have high stakes. For instance, observations play an important role in evaluating teachers or their teaching. Similarly, observations may be a key source of data in assessing the impact of a new policy or instructional innovation. Typically, researchers have used generalizability theory (or a g-study) to determine how many observations are needed to achieve acceptable reliability. Here we outline the results of prior studies.

The first study of interest considered the Mathematical Quality of Instruction (MQI) instrument (Hill et al. 2008) and the number of lessons and raters required to achieve sufficient reliability for capturing teaching (Hill et al. 2012). The g-study was conducted using a sample of 8 middle school mathematics teachers, each who were observed four times. The scoring was completed by a team of 10 graduate students and teachers who underwent a two-day training. Each video was scored by four raters.

Ultimately, the research team found it was needed to have four raters scoring four lessons to reach a standard of 80% reliability, which was still below the high standard of 90% reliability for policy decisions. Extrapolations from the D-study suggested that around five to 12 observations would be required from a single observer for the same reliability. Nonetheless, the authors did make recommendations on observations for other purposes. It was noted that there were diminishing returns in adding additional observations or raters. Given this, it was suggested that three lessons with two raters was the optimal combination for research purposes, taking in the additional costs of more lessons and more raters (Hill et al. 2012).

Another study, which focused on the generalizability of observations from the Trends in International Math and Science Study (TIMSS), found that five to six observations with four raters were needed to provide a satisfactory level (80%) of reliability for decision-making (Newton 2010). However, given the cost of multiple observations, the authors suggested that four observations with four raters would be an acceptable compromise. Notably, the researchers found that there was more variation (requiring more observations) for elementary school as compared to middle school classrooms.

In a third study, the International Comparative Analysis of Learning and Teaching (ICATL) observation protocol was used to study classroom teaching in the Netherlands (van der Lans et al. 2016). This study focused on classrooms across

disciplinary areas, with 22% of the sampled classrooms in mathematics. To achieve the cutoff set for formative feedback (70% reliability), a total of three observations with a single rater were needed. To achieve the high cutoff needed for evaluation decisions (90% reliability), more than 10 observations would be needed.

A final study, involving the Toolkit for Assessing Mathematics Instruction (TAMI-OP), found similar results (Weston et al. 2020). This study differed from others, because it focused on undergraduate mathematics. The TAMI-OP was based on the Classroom Observation Protocol for Undergraduate STEM (COPUS; M. K. Smith et al. 2013), with the time-segmented features of the Teaching Dimensions Observation Protocol (TDOP; Hora 2015). In this study, it was found that 11 observations over a single semester with a single rater were needed to have a reliable measure at 80%.

Summarizing across these studies, it is clear that a large number of observations and/or raters are needed to achieve sufficient reliability to use classroom observations for high-stakes policy decisions or teacher evaluations. Relaxing the standards to use data for formative feedback, this research suggests that at least three (Hill et al. 2012), but ideally four to six observations would take place (Hill et al. 2008; Newton 2010). As prior work highlights, the number of observations required depends on the tool, the context, and the purpose of the observations. Regardless of the specifics, it is clear that the one to two observations used in many research studies does not meet these recommended guidelines. We now turn our attention to EQUIP, which differs from previously studied tools in important ways.

EQUIP observation tool

Equity QUantified in Participation (EQUIP) is a classroom observation tool that was designed to capture patterns of equity and inequity in STEM classroom discourse (Reinholz & Shah 2018). EQUIP is also freely available as a web app (https://www.equip.ninja), which streamlines the process of performing a classroom observation with the EQUIP protocol. EQUIP was designed to answer questions such as: What proportion of high-level questions are asked to women? Or, in what ways do Indigenous students contribute to the classroom discussion? Although no external tool can objectively define equity in a classroom (Garcia et al. 2018), EQUIP aims to provide data analytics that can identify sources of *inequity*, and can also be given to instructors to support professional development (Reinholz et al. 2019a, b).

EQUIP was developed through an extensive literature review, consultations with experts in educational equity, and extensive preliminary analyses from a team of five raters (Reinholz & Shah, 2018). Since its initial development, EQUIP has been used to study hundreds of classrooms across a variety of studies in which multiple raters achieved high levels of inter-rater reliability (Reinholz & Wilhelm, forthcoming; Reinholz et. al. in press). EQUIP is flexible insofar that there are a variety of use cases which vary across studies. For instance, EQUIP can be used real-time by an observer in a classroom (using the EQUIP app), it can be applied to classroom videos without transcripts, or it can be used to code transcripts of classroom discourse.

The basic unit of analysis in EQUIP is a participation *sequence*. A sequence constitutes a segment of talk by one student (and possibly the teacher) that is uninterrupted by another student. Thus, each sequence corresponds to only one student, which allows all coded participation to be assigned to particular students. For each coded sequence, features of the participation are coded to capture the quality of a student's contribution. By adding demographic information about students in a class, EQUIP generates analytics for social marker groups. The demographic categories used are customizable, and prior research has focused on markers such as: race, gender, first-generation status, dis/ability, and socioeconomic status. The choice to attach sequences to particular students allows for disaggregation of participation, but as a result, EQUIP does not capture student–student interactions in discourse, unless a user specifically customizes for it.

Like the demographic categories, EQUIP features customizable codes. Because EQUIP is customizable and focuses on quantifiable aspects of participated coded at the individual level, these codes may vary from study to study, but by default, EQUIP looks at features of participation such as: the student talk length, the student talk type, a teacher's questions, how a student is called on, and how a teacher responds to a student's talk. Furthermore, EQUIP differs from a typical protocol, for instance, which relies on rubrics and affords opportunities to analyze patterns within and across the individual codes. Instead, EQUIP will capture the number of instances and characteristics of particular codable events related to participation. In general, the codes generated by EQUIP are categorical, although it is possible that in some cases the codes may have meaning as an ordinal scale. Because of this unique design feature, analyses of data generated by EQUIP may utilize Chi-Squared, Fisher's Exact Test, Cramer's V, and other less commonly used statistical measures.

Methods

Data sources

The data analyzed for this study were drawn from a prior research study that focused on equitable teaching in inquiry-oriented undergraduate mathematics (Smith et al. 2019, Author et al., in press). In that study, EQUIP was used to analyze the teaching of 42 instructors who were using inquiry-based curriculum that were designed for one of three courses: Linear Algebra (LA), Abstract Algebra (AA), or Differential Equations (DE). Instructors in the project received ongoing professional development through summer workshops and online working groups. The purpose of the professional development was to help instructors attend to the four key components of inquiry-oriented-instruction: generating student ways of reasoning, building on student contributions, developing shared understandings, and connecting to standard mathematical language and notation (Kuster et al. 2018).

For each instructor in the study, two units of teaching were observed and recorded. Here we focus on the second unit of teaching, which was recorded 8–10 weeks into the semester. We choose this focus because the later unit was best able to capture established teaching practices and classroom norms. Teachers were observed over

multiple, subsequent class sessions to ensure that the unit could be captured in its entirety (mean = 149 min, SD = 61 min). The EQUIP protocol was used to record these video records of classroom interactions, and in the process, the research team generated partial transcripts that could be referred back to later. The dataset was coded by a team of three raters, who achieved a high-level of interrater reliability (over 0.8 for Krippendorf's alpha on each dimension). After inter-rater reliability was achieved on a sufficient subset of the whole dataset (over 20%), the rest of the class sessions were coded individually. Because we did not have multiple raters for all classrooms in the dataset, in our analyses below, we focus only on a single rater.

Given that the length of the observed units differed between instructors, we created a standardized unit of a single 50-min class session, which mirrors the standard class session for a class meeting 3 times per week. Once standardized, we calculated the average number of contributions per 50-min unit (mean = 23 contributions, SD = 17.7). Based on prior studies, we assumed that at least three observations would be need for stability, so in our analyses below we only included classrooms for which there were more than 69 contributions total ($N = 20$ classrooms met this criterion).

Of the 42 instructors in the larger study, prior work identified a *different* subset of 20 instructors for which suitable measures of student gender and assessment data were available (Reinholz et al. in press). A finding of that study was that women's average levels of participation significantly predicted gendered performance differences in outcomes. In other words, EQUIP data were useful in identifying sources of gender inequity in student outcomes and linking them to classroom participation. This is consistent with broader literature that highlights the connections between student participation and performance (e.g., Banes et al. 2019).

Codes

As mentioned previously, EQUIP is fully customizable. In Table 1, we provide a list of the codes that were used in the current study (Reinholz et. al. in press). These particular codes were customized to capture features of inquiry-oriented instruction.

In addition to this set of codes, the study also tracked the gender of participating students. While other studies with EQUIP have focused on student race, those data were missing in the present study, so they could not be included in our analyses. A thorough explanation of typical EQUIP codes and their theoretical grounding is described in depth elsewhere (Reinholz & Shah, 2018; Author et al. in press).

In Table 2, we provide a general set of descriptive statistics for the dimensions coded in the study.

G-theory

G-theory is a statistical method used to quantify the stability of behavioral measurement (Brennan 2001). There are two components in g-theory: (a) generalizability study (g-study), and (b) decision study (d-study). G-study helps quantifying the variance of measurements. D-study helps identify the optimal number of classroom

Table 1 EQUIP codes used to capture inquiry-oriented instructional discourse

Code	Subcode	Definition
Instructor solicitation method	Called on	Instructor calls on a student
	Not called on	A student interjects without being called on by instructor
Instructor solicitation type	Why	Instructor asks student to explain/justify their reasoning
	How	Instructor asks for a student's solution method
	What	Instructor asks a student to read part of a problem, recall a fact, or give a numerical/verbal answer
	Other	Instructor asks a general question (e.g., "What did you think?")
	N/A	Instructor does not ask the student a question
Instructor response	Elaborate	Instructor expands on or formalizes the student's idea
	Revoice	Instructor repeats student contribution
	Evaluate	Instructor explicitly says the student is correct/incorrect
	Follow-up	Instructor asks a follow-up question and a new student responds
	N/A	Instructor does not respond to the student's contribution
Student talk length	21 + words	Student speaks 21 + words consecutively
	5–20 words	Student speaks 5–20 words consecutively
	1–4 words	Student speaks 1–4 words consecutively
Student talk type	Why	Student explains/justifies their reasoning
	How	Student describes solution method
	What	Student reads part of the problem, recalls a fact, or gives a numerical/verbal answer to a problem
	Other	Student asks a question or says something nonmathematical

observations to maximize reliability by analyzing the generalizability coefficient ($\boldsymbol{E}\rho 2$) and index of dependability (Φ) (Brennan 2001).

Since the data in this study were only coded by one rater, we conducted a one-facet g-study for each code to determine the stability for the respective code, where the codes were treated as scores and the lessons for each class were treated as facets.

The data were transformed to fit the ordinal or continuous requirements of g-theory. Particularly, since student talk length had a natural ordinal order, it was transformed into a continuous variable (0: *1–4 Words*, 1: *5–20 Words*, 2: *21 + Words*). Binary data (*Instructor Solicitation Method* and *Gender* – gender data were a binary as a limitation of secondary analysis), were coded as binary (e.g., 0: *Called On* or 1: *Not Called On* for *Instructor Solicitation Method*). Finally, for non-binary categorical codes, we treated each subcode as a binary code. For example, the *Student Talk Type* code was broken down into *Student Talk Type How* (0: *Not How*, 1: *How*), *Student Talk Type What* (0: *Not What*, 1: *What*), *Student Talk Type Why* (0: *Not Why*, 1: *Why*), and *Student Talk Type Other* (0: *Not Other*, 1: *Other*). This was repeated for *Instructor Solicitation Type* and

Table 2 Descriptive statistics of the combined classrooms in the study

Input	Description	Count	Proportion (%)
Instructor solicitation method	Called on	588	(22)
	Not Called on	2073	(78)
Instructor solicitation type	How	69	(3)
	Other	384	(14)
	What	919	(35)
	Why	213	(8)
	N/A	1074	(40)
Instructor response	Elaborate	593	(22)
	Evaluation	137	(5)
	Follow up	391	(15)
	Revoice	384	(14)
	N/A	1152	(43)
Student talk length	1–4 Words	645	(24)
	5–20 Words	1414	(53)
	21 + Words	601	(23)
Student talk type	How	130	(5)
	What	1633	(61)
	Why	248	(9)
	Other	648	(24)
Gender	Man	1834	(71)
	Woman	766	(29)

Instructor Response. After all d-studies were completed, we plotted the generalizability coefficient across the d-study sample sizes. We particularly looked for diminishing returns of increased sample size (Huebner and Lucht 2019). All g-theory analyses were conducted using the gtheory package in R (Moore 2016). For more details on how to conduct g-theory in R see Huebner and Lucht (2019).

Simulation

We conducted a simulation study to complement the g-study. We used a simulation in addition to a g-theory analysis for multiple reasons. First, an important assumption of a g-study is that there is one continuous or ordinal behavioral measurement that is treated as the score (e.g., a standardized quantitative measure calculated using a rubric) that summarizes a classroom observation. However, EQUIP does not provide single outcome variable, but rather emphasizes multiple dimensions of student participation which are not always continuous or ordinal (many are categorical). Second, and related to variable type, variable components in g-theory will traditionally treat variables as continuous (Ark 2015), but there are both continuous and categorical data in EQUIP. Finally, there is no "ideal" measure, so whether one particular distribution of participation is preferable to another is a matter of interpretation

within the local context (e.g., a *why* statement isn't necessarily always better than a *how* statements for *Student Talk Type)*. Our analyses needed to be sensitive to these types of issues.

For this reason, we needed a statistical measure that allowed for mixed variable types (both in the score output and variable inputs) as well as individual and aggregate analyses of the codes. Particularly, in addition to using a g-study analysis that required us to separate all codes into subcodes, we used a simulation that aims to mimic data collection process for varying number of classroom observations (e.g., one observation, two observations, and so on) to identify the optimal number of observations that accurately captures the teaching practices measured by EQUIP. Simulations were especially useful in this study since they were situated in the data and were not dependent on meeting statistical assumptions. This algorithm we used for simulations is shown in Table 3. By using both g-theory and the simulations, we are able to continue conversations about to relate to the existing literature that uses g-study and provide an additional method to support or challenge the results of a g-theory analysis.

We began by randomly sampling consecutive sequences of size S for a varying number of sampled observations (from 1 to i) from the complete observed data (Line 3). In this simulation, we only allowed for consecutive sequences to account for dependency between time-adjacent sequences. Although the algorithm allows the user to select the number length of S, S was 21 for all classrooms in this data because this was the average number of sequences across inquiry-oriented undergraduate mathematics classrooms from the larger sample. In all classrooms, we allowed up to $i=20$ observations. We then calculated the Cramer's V for each code by comparing the randomly sampled data from Line 2 to the complete observed data (Line 5). Cramer's V measures the association between categorical data, where Cramer's V closer to zero implies that there is a weak effect. In this simulation, Cramer's V is a measure of accuracy where a small Cramer's V implies that the randomly sampled data and complete sampled data are similar whereas a large Cramer's V implies that there is a difference between both data. Finally, after $j=500$ iterations, the combined weighted average Cramer's V (Line 9) for each sample observation O_i and iteration j was:

Table 3 Simulation algorithm

Line	Code
1	**For** each number of sampled observations from 1 to i
2	**For** each iteration from 1 to j
3	Randomly sample S consecutive sequences from the *complete observed data*, with replacement, to create the *randomly sampled data*
4	**For** each code k
5	Calculate the *Cramer's V* comparing the *randomly sampled data* to the *complete observed data* for this code ($C_{i,j,k}$)
6	**End**
8	**End**
9	Calculate the *combined weighted average* Cramer's V across the codes k for each sample observation O_i using inverse variance weights

$$\frac{\sum^{k}\left(w_{i,k}\cdot C_{i,j,k}\right)}{\sum^{k} w_{i,k}}$$

where $w_{i,k}=1/\sigma^{2}_{i,k}$ and $\sigma^{2}_{i,k}$ is the variance for code k generated using the sampling distribution in Line 5 for the sample observation O_i across all iterations. In other words, we found the average Cramer's V for each code k in iteration j, then used inverse variance weights to find the combined average Cramer's V across all the codes for each sample observation O_i in iteration j, and used this to create a sampling distribution of the weighted average Cramer's V across all iterations j for each sample observation O_i.

For example, to randomly simulate two classroom observations ($i=2$) in the 15FA01 data, with an average of 21 sequences per observation ($S=21$) and six codes, we began by randomly sampling 42 consecutive sequences from the complete observed data. For each of the six codes, we calculated the Cramer's V comparing the 42 randomly sampled consecutive sequences to the original data for the eight codes, found the combined weighted averaged of the Cramer's V across the $j=500$ randomly sampled observations, and created a sampling distribution of the Cramer's V for each code as well as for the combined weighted average to create inferences about the generalizability of O_i observations.

There are a few items worth noting in this simulation. First, Cramer's V measures the effect size in categorical data across different groups that is based on Pearson's chi-square statistic. Unlike the phi statistics, Cramer's V measures the association between two categorical variables when the contingency table is larger than 2×2 and accounts for small sample sizes. The degrees of freedom for all the Cramer's V is *min(r-1,c-1)*, where r and c represent the number of rows and columns on a contingency table, respectively. In this simulation, the contingency table included the categories of a code (varied by code) and the two groups (the complete observed data and randomly sampled data). Since there were only two groups, regardless of the code, the degrees of freedom for each Cramer' V was $df=1$, allowing us to average across and within the different codes within each randomly sampled data.

Second, unlike g-study analyses, the simulation afforded opportunities for us to create inferences about the Cramer's V within the observations, such as creating a sampling distribution of the mean Cramer's V for each code or across codes, to guide us in identifying the optimal number of classroom observations. In this study, we considered multiple criteria to determine the optimal number of classroom observations: (a) classroom observations that have an average Cramer's V below 0.10, (b) classroom observations that have at least 95% of the simulated Cramer's V below the 0.10 threshold, and (c) have at least three classroom observations. The first criterion was selected in alignment with Cohen's (1988) recommendations for interpreting the Cramer's V, where Cramer's V below 0.10 imply that there is no trivial difference between two groups. The second criterion extends the first criteria by ensuring that the majority if the Cramer's V from the randomly sampled observations are below the 0.10 threshold level. The third criteria builds on the current literature that suggests that at least three classroom observations are required to find the optimal number of classroom observations (Hill et al. 2012).

Finally, given that not all classroom observations had the same number of sequences, we focused on the analysis of *consecutive sequences* in our simulations. In this way, the sequences we used either represented multiple sequences from the same class session, or sequences from time-adjacent class sessions. In our algorithm, we always sample subsequent sequences, to account for the time-dependence of observations. We recognize multiple sources of variation between observations. On one hand, we expect that observations would look different over a semester or year. In our study, we have subsequent observations in a single unit, so we may not fully capture this variation. We also except that there is random variation between any two given days, given changes in teaching practices and different types of lessons. This is the type of variation we aimed to uncover through the results of our simulations.

Results and discussion

Our results provide multiple analyses to characterize the stability of observations with EQUIP. The results section is divided into three parts. First, we provide an aggregate analysis of Cramer's V across all 20 classrooms in the sample. This provides an overall picture of the stability of observations. Second, we disaggregate by specific type of code, to show the stability of different constructs within EQUIP. Third, we present the results of the g-study, which is disaggregated by each level of each code (a further level of disaggregation).

Simulations

Stability by class

Our first set of simulations focused on the stability of EQUIP observations across classes. These analyses were completed using an average weighted Cramer's V, shown in Fig. 1. The inverse variance weights were especially useful in this simulation since they minimize the variance of the weighted averages by placing a heavier weight on codes with a smaller variance.

As expected, we see a decrease in the Cramer's V and the 95% lower bound as the number of observations increases for all 20 classes. The average Cramer's V and 95% lower bound for all the courses was below the 0.10 threshold after one classroom observation. The 95% lower bound varied slightly by classroom. Particularly, three classes had a 95% lower bound at or below 0.10 after three observations (DE1 DE8, LA2), and the remaining classes had a 95% lower bound at or below 0.10 after four classroom observations. In other words, all classes had an average Cramer's V and 95% lower bound after at least four classroom observations. These results indicate that—at least for the classes in this sample—they did not differ greatly in terms of the overall number of observations required to reach some level of stability with Cramer's V.

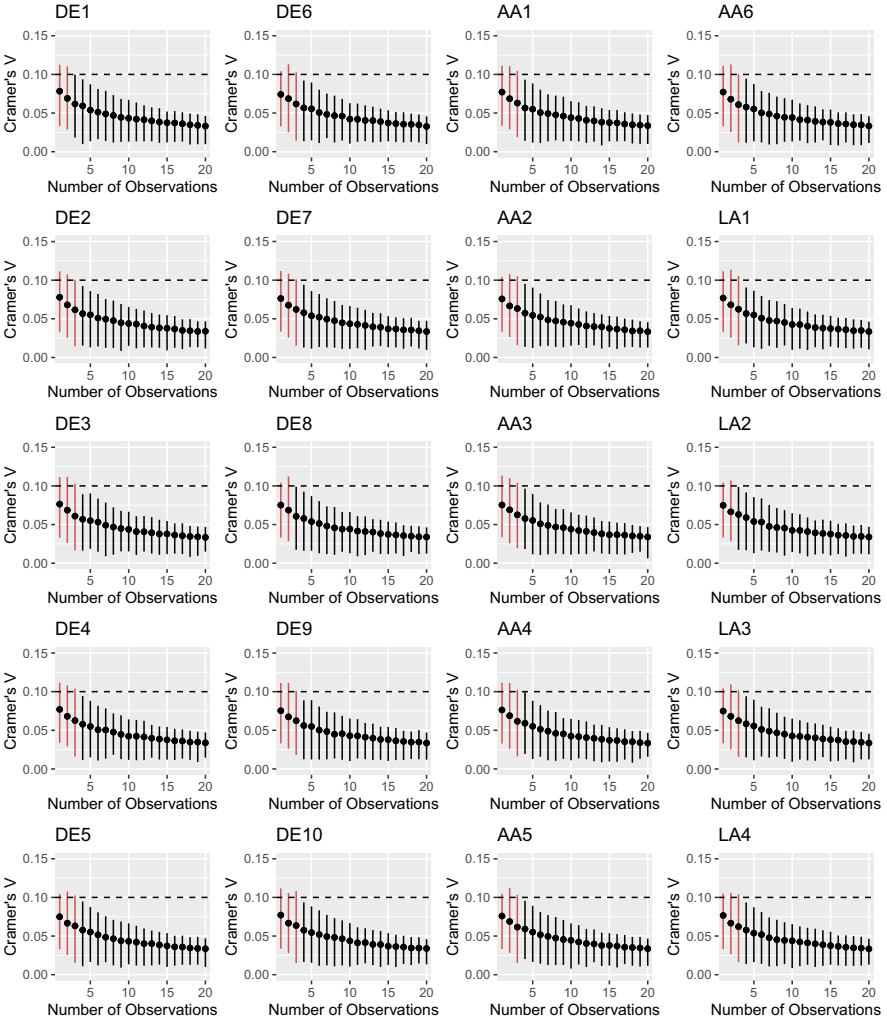

Fig. 1 Forest plots of the combined Cramer's V for up to 20 observations across classes. The vertical lines represent the bottom 95% bounds for the Cramer's V for each randomly sampled observation, up to twenty. The 95% lower bounds that cross the 0.10 threshold are in red. Similarly, the average Cramer's V within each randomly sampled observation are shown in points, where the red points are averages above 0.10

Stability by code

Our next set of analyses focused on the stability of each individual code within EQUIP. To illustrate our results, we provide an in-depth analysis of the overarching trends across one class, DE1, shown in Fig. 2. This particular class was chosen as it was representative of the general trends we found. We provide a summary of the

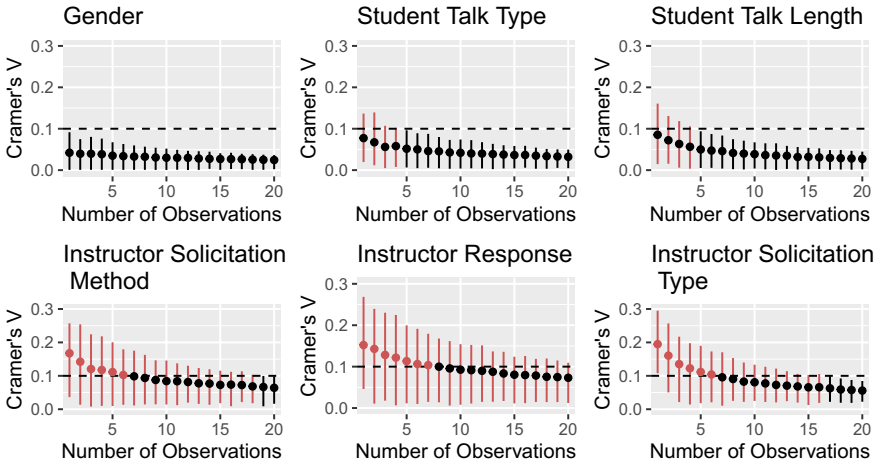

Fig. 2 Modified forest plots of Cramer's V for up to 20 observations for the codes in DE1 (S = 21)

Table 4 Number of observations to achieve an average Cramer's V (Avg CV) and 95% lower bound (95% LB) at or below 0.10 in all classes

	Gender	Student talk type	Student talk length	Instructor solicitation method	Instructor response	Solicitation method
Avg CV	1	1	1	7	9	7
95% LB	1	6	5	17	20+	19

results for all classes in Table 4, and we will provide visuals for the entire dataset in our supplementary materials. The 95% lower bound of the combined Cramer's V within each of the 20 randomly simulated classroom observations are shown using vertical lines. Lower bounds that cross the 0.10 threshold are shown in red and those at or below the 0.10 are shown in black. The averages within each number of observations are also shown, where averages above the 0.10 are shown in a red point and those at or below 0.10 are shown in a black point.

In Fig. 2, we see that all the codes had an average Cramer's V at or below the 0.10 threshold after at most eight observations, with *Student Talk Length*, *Student Talk Type*, and *Gender* having an average Cramer's V below 0.10 after one observation. In terms of the 95% lower bound, *Gender*, *Student Talk type*, and *Student Talk Length* were at or below the 0.10 threshold after five classroom observations. The *Instructor Response*, *Instructor Solicitation Type*, and *Instructor Solicitation Method* codes varied more across the number of observations, requiring more than 20, at least 17, and at least 19 classroom observations, respectively, to show a 95% lower bound at or below the 0.10 threshold. In other words, participation as characterized by *Instructor Response*, *Instructor Solicitation Type*, and *Instructor Solicitation Method* varied more than the other codes across DE1. We interpret

this to mean that the ways in which instructors used evaluation, solicitation types, and solicitation methods differed more across particular lessons, whereas the gendered distributions of participation and the lengths and types of student talk were more stable. It could be that while a teacher may have used a variety of new different strategies across lessons, once a classroom culture becomes established, student participation patterns are more ingrained. In addition, the *Instructor Response*, *Instructor Solicitation Type*, and *Instructor Solicitation Method* codes had a large number of categories, which inherently adds more variance. One possible way to address this would be to condense the codes into fewer categories.

These results are shown to highlight how researchers can consider individual codes to select the optimal number of classroom observations. For example, if a researcher is interested in stabilizing the *Gender* code (which would be one signifier of participatory equity) as measured by the average Cramer's V and 95% lower bound, then only a few observations may be needed. However, if the researcher is interested in getting an accurate and efficient interpretation of teacher evaluation, they may need many more observations.

In terms of code stability across all 20 classes, Table 4 shows the number of observations needed to achieve an average Cramer's V and 95% lower bound at or below 0.10 for all the codes. The *Gender* code is the most stable since it achieved an average Cramer's V and 95% lower bound at or below 0.10 in all classes after one classroom observation. The *Student Talk Type* achieved an average Cramer's V and 95% lower bound after six classroom observations. Additionally, *Student Talk Length* achieved an average Cramer's V and 95% lower bound after five observations. Although all the other codes (*Instructor Response*, *Instructor Solicitation Type*, and *Instructor Solicitation Method*) had an average Cramer's V below 0.10 after at most ten observations, they also all had a 95% lower bound after at least 20 classroom observations. This may be telling of the variety in classroom participation as measured by these codes.

G-theory

To supplement the simulations study, we also used g-theory. After all g- and d-studies were conducted, we plotted the generalizability coefficients in relation to the d-study sample sizes. The plot for all the codes is shown in Fig. 3. As a reminder, *Student Talk Type*, *Instructor Response*, and *Instructor Solicitation Type* were transformed to treat the subcodes as binary codes. The *Student Talk Length*, *Gender*, and *Instructor Solicitation Method* codes were already ordinal or binary.

Overall, we see that increasing the number of observations increases the generalizability code. However, the generalizability coefficient and rate at which they achieved high generalizability coefficient varied by code. For example, the *Gender* code achieved a relatively high generalizability coefficient and diminishing returns after about five observations. Although the simulation suggested at least one observation (a lower number), these results were still consistent across the simulation and g-study, as *Gender* appeared to be the most stable code in each case.

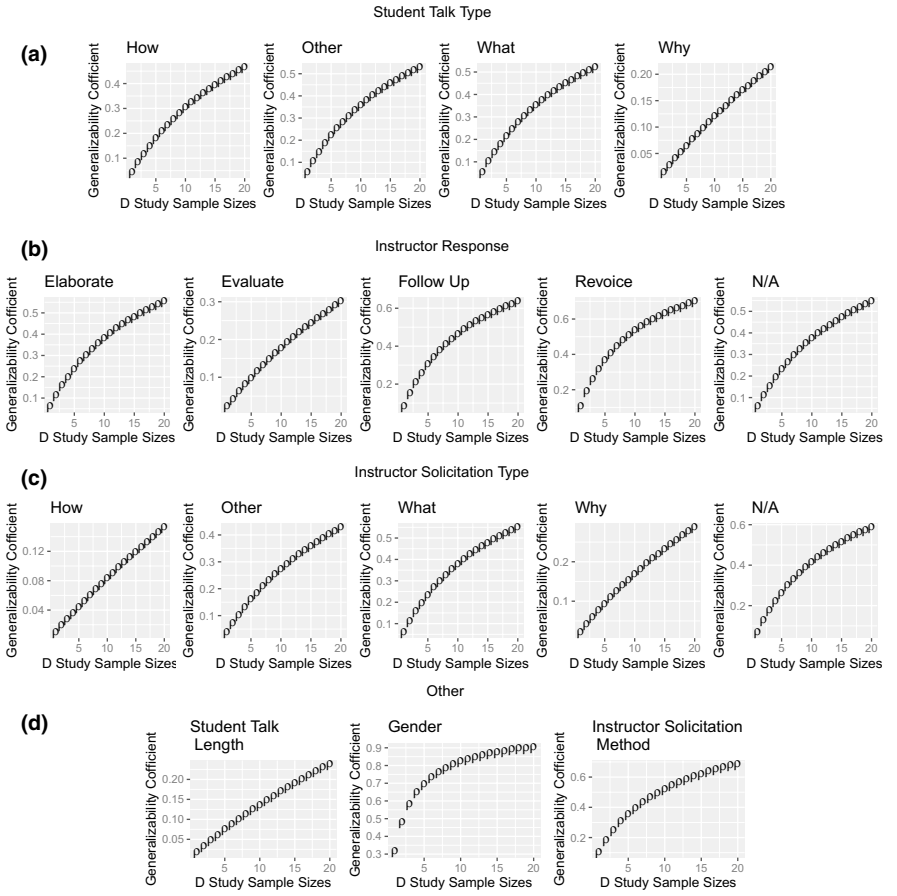

Fig. 3 The reliability of the student talk type, instructor response, instructor solicitation type, student talk length, gender, and instructor solicitation method

One interpretation of this pattern is that the distribution of participation by *Gender* remained consistent across most classrooms. On the other hand, the majority of the other codes (e.g., *Instructor Solicitation Method*, *Student Talk Length*, *Student Talk Type*, *Instructor Solicitation Type*, *Instructor Response*) did not appear to achieve diminishing returns after about 10 observations. It is also worth noting that some codes and subcodes that had relatively small occurrences (e.g., *Instructor Solicitation Type*, *Student Talk Type Why* and *Other*, and *Instructor Response Elaborate*) generally had lower generalizability coefficients. This may be attributed to the variance of the codes and subcodes. Nonetheless, it suggests that variables with high variance and possibly low frequencies require more than 10 observations to achieve diminishing returns.

Conclusion

Our study used both a g-study and simulations to study in-depth the stability of classroom observations with the EQUIP tool. Our findings have a number of insights for use of the EQUIP tool and observations more generally. On the whole, our results showed that the overall Cramer's V for each class across the simulation study tended to achieve stability very quickly, which meant that we did not find evidence of large amounts of variance *between classes*.

However, when we looked at individual EQUIP codes, we did find a fair deal of variation. Of particular interest was that the *Gender* code was amongst the most stable in both the simulations study and the g-study. This suggests that when using EQUIP to capture overall patterns of inequitable participation, just a few observations may be sufficient. In our simulations study, we also found that *Student Talk Type* and *Student Talk Length* became stable after just one observation on average, or 5–6 observations for stability with the 95% confidence interval. This contrasted the instructor codes which varied much more. *Instructor Response*, *Instructor Solicitation Type*, and *Instructor Solicitation Method* only achieved diminishing returns at closer to 20 observations. One possible interpretation is that while instructor moves may vary more between different lessons, once patterns of student participation become entrenched, they are relatively stable. Another possible interpretation relates to the inherent variance of codes with many levels. For instance, there were many levels for *Instructor Response* and some of them occurred rather infrequently, which could result in large variations between lessons.

To interpret these results, we reiterate the context of inquiry-oriented teaching in undergraduate mathematics. This is a context with free-flowing discussions and relatively high-levels of classroom discourse, which may result in greater variation across lessons. We suspect that a more structured setting (e.g., a traditional Initiate-Response-Evaluate-focused classroom in middle school mathematics) might have much less variation across lessons. We also note that our observations were sampled from subsequent lessons, and if the observations were taken from different parts of the semester there would likely be more variation. Thus, when generalizing the results into other contexts of observation, a reader should carefully account for potentially sources of variation and how they would compare to this benchmark study. In addition, a user must consider their goals of the observations. While a few observations are likely sufficient for professional development (at least along some measures), a much larger number would be desirable for policymaking. We also recommend that users can reduce the number of levels for each code as a strategy to reduce variation in observations, as this would limit the occurrence of codes that arise with very low frequency and thus add greatly to variation.

We also remark on the methodological contributions of this work. In our study we used both a g-study and simulations with Cramer's V. On the whole we found that the results were consistent (at least in terms of which codes were more or less stable), but the g-study predicted a higher number of observations needed to achieve stability. We suspect that one reason for this was that the g-study required us to split up variables that were naturally categorical, which resulted in a large number

of subcodes, many of which had a small number of occurences. The simulations approach was powerful because it allowed us to quantify the stability of mixed behaviorial measurements, rather than relying on continuous or ordinal measurements, as was typical for g-theory. Our work provides a useful starting point for others who wish to study the stability of mixed behaviorial measurements like those used in the EQUIP tool. We recommend that using g-theory, simulations, and other potential approaches in conjunction can give a more holistic view.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s43545-021-00190-x.

Funding This material was partially supported by the National Science Foundation under Grant No. 1943146.

Data availability The data from this study are not publicly available. The EQUIP observation tool is freely available at https://www.equip.ninja.

Code availability There is no code available.

Declarations

Conflict of interest The authors declare they have no competing interests.

Ethical approval All research was completed with permission from the Insitutional Review Board at San Diego State University.

References

Ark TK (2015) *Ordinal generalizability theory using an underlying latent variable framework* [University of British Columbia]. https://doi.org/10.14288/1.0166304

Banes LC, Restani RM, Ambrose RC, Martin HA, Bayley R (2019) Relating performance on written assessments to features of mathematics discussion. Int J Sci Math Educ. https://doi.org/10.1007/s10763-019-10029-w

Brennan RL (2001) Generalizability theory. Springer, New York

Ernest JB, Reinholz DL, Shah N (2019) Hidden competence: Women's mathematical participation in public and private classroom spaces. Edu Studies Math, 102(2):153–172. https://doi.org/10.1007/s10649-019-09910-w

Garcia NM, López N, Vélez VN (2018) QuantCrit: rectifying quantitative methods through critical race theory. Race Ethn Educ 21(2):149–157. https://doi.org/10.1080/13613324.2017.1377675

Hill HC, Blunk ML, Charalambos CY, Lewis JM, Phelps GC, Sleep L, Ball DL (2008) Mathematical knowledge for teaching and the mathematical quality of instruction: an exploratory study. Cogn Instr 26:430–511

Hill HC, Charalambous CY, Kraft MA (2012) When rater reliability is not enough: teacher observation systems and a case for the generalizability study. Educ Res. https://doi.org/10.3102/0013189X12437203

Hora MT (2015) Toward a descriptive science of teaching: how the TDOP illuminates the multidimensional nature of active learning in postsecondary classrooms. Sci Educ 99(5):783–818. https://doi.org/10.1002/sce.21175

Huebner A, Lucht M (2019) Generalizability Theory in R. Pract Assessment Res Evaluat. https://doi.org/10.7275/5065-gc10

Kuster G, Johnson E, Keene K, Andrews-Larson C (2018) Inquiry-oriented instruction: a conceptualization of the instructional principles. Primus 28(1):13–30. https://doi.org/10.1080/10511970.2017.1338807

Moore CT (2016) *Package 'gtheory'*. https://cran.r-project.org/web/packages/gtheory/gtheory.pdf

Newton XA (2010) Developing indicators of classroom practice to evaluate the impact of district mathematics reform initiative: A generalizability analysis. Stud Educ Eval 36(1):1–13. https://doi.org/10.1016/j.stueduc.2010.10.002

Reinholz DL, Shah N (2018) Equity analytics: A methodological approach for quantifying participation patterns in mathematics classroom discourse. J Res Math Edu 49(2):140–177

Reinholz DL, Stone-Johnstone A, Shah N (2019a) Walking the walk: using classroom analytics to support instructors to address implicit bias in teaching. Int J Acad Develop. https://doi.org/10.1080/1360144X.2019.1692211

Reinholz DL, Bradfield K, Apkarian N (2019) Using analytics to support instructor reflection on student participation in a discourse-focused undergraduate mathematics classroom. Int J Res Undergrad Math Edu 5(1):56–74. https://doi.org/10.1007/s40753-019-00084-7

Reinholz DL, Johnson E, Andrews-Larson C, Stone-Johnstone A, Smith J, Mullins B, Fortune N, Keene K, Shah N (in press) Is active learning equitable without equitable participation?: Women's participation predicts gender inequities in student performance. J Res Math Edu

Reinholz DL, Wilhelm AG (under review) Race-gender D/discourses in mathematics education: (Re)-producing inequitable participation patterns across a diverse, instructionally-advanced district

Shah N, Christensen JA, Ortiz NA, Nguyen A, Byun S, Stroupe D, Reinholz DL (2020) Racial hierarchy and masculine space: Participatory in/equity in computational physics classrooms. Com Sci Edu 1–25. https://doi.org/10.1080/08993408.2020.1805285

Smith MK, Jones FHM, Gilbert SL, Wieman CE (2013) The classroom observation protocol for undergraduate STEM (COPUS): a new instrument to characterize university STEM classroom practices. CBE-Life Sciences Education 12(4):618–627. https://doi.org/10.1187/cbe.13-08-0154

Smith J, Andrews-Larson C, Reinholz DL, Stone-Johnstone A, Mullins B (2019) Examined inquiry-oriented instructional moves with an eye toward gender equity. Proceedings of the 2019 conference on research in undergraduate mathematics education

van der Lans RM, van de Grift WJCM, van Veen K, Fokkens-Bruinsma M (2016) Once is not enough: establishing reliability criteria for feedback and evaluation decisions based on classroom observations. Stud Educ Eval 50:88–95. https://doi.org/10.1016/j.stueduc.2016.08.001

Weston TJ, Hayward CN and Laursen SL (2020) When seeing is believing: generalizability and decision studies for observational data in evaluation and research on teaching. *American Journal of Evaluation*