Int J of Sci and Math Educ (2015) 13:389–403
DOI 10.1007/s10763-014-9610-5

Where Low and High Inference Data Converge: Validation of CLASS Assessment of Mathematics Instruction Using Mobile Eye Tracking with Expert and Novice Teachers

Kai S. Cortina · Kevin F. Miller · Ryan McKenzie · Alanna Epstein

Received: 27 May 2014 / Accepted: 11 December 2014 / Published online: 21 February 2015

Abstract Classroom observation research and research on teacher expertise are similar in their reliance on observational data with high-inference procedure to assess the quality of instruction. Expertise research usually uses low-inference measures like eye tracking to identify qualitative difference between expert and novice behaviors and cognition. In this study, we used mobile eye-tracking technology to create a low inference quality indicator for the comparison of experienced and student teachers. The distribution of visual fixations on students was measured using Gini coefficients based on the observation of van den Bogert, van Bruggen, Kostons, and Jochems (*Teacher and Teacher Education*, *37*, 208–216, 2014) that expert teachers show better classroom monitoring. Results confirm that student teachers have a higher Gini coefficient than experienced teachers indicating weaker classroom monitoring. However, the Gini coefficient did not correlate in the predicted way with trained observer coding of video footage of the same classrooms using the Classroom Assessment Scoring System (CLASS) (Pianta, Hamre, Haynes, Mintz, & La Paro, 2007) although the mean differences in behavioral management were higher for the experienced teachers as expected. The CLASS dimension Quality of Feedback was significantly related to the Gini coefficient as an interaction with expertise: Only for novice teachers that a high quality of feedback was negatively associated with monitoring of the classroom.

Keywords Mobile eye tracking · Classroom observational assessment · CLASS · Expert–novice paradigm · Quality of feedback

K. S. Cortina (✉) · K. F. Miller · R. McKenzie · A. Epstein
Department of Psychology, University of Michigan, 530 Church St, Ann Arbor, MI 4809-1043, USA
e-mail: schnabel@umich.edu

Theoretical Background

The role of the classroom teacher and the characteristics of effective teaching for successful learning in school have remained key topic in empirical educational research over almost five decades. Researchers have long debated not only what it is that makes a teacher effective but also how these effects should be measured and how teacher training could benefit from this line of empirical research (Campbell, Kyriakides, Muijs & Robinson, 2003; Muijs & Reynolds, 2001). With the paradigm shift away from looking solely at the teacher and his or her personality and fitness for the job towards a more interactive process–product paradigm in the early 1970 came a focus on observational assessments of classroom quality starting with the work by Flanders (1970) and the seminal work by Brophy & Good (1974, 1986). A multidimensional perspective characterizes the teacher effectiveness paradigm that emerged in the mid-1980s (Bromme, 2001; Seidel & Shavelson, 2007). Classrooms were now seen as complex social interactions that differed in organization, climate, pace, etc. It became obvious that there is no one-fits-all-needs optimal teaching practice. Balancing cognitive and social aspects of learning, as well as the trade-off between excellence of some students, on the one hand, and learning progress of the entire learning group, on the other hand, makes it necessary for a teacher to be explicit about the learning goals and to choose teaching strategies accordingly. The argument implies that good instruction is something that is based on key techniques that a teacher can acquire through adequate teacher training and professional development. Two distinct research traditions have evolved in the field since: teacher expertise research (Bromme, 2001) and research on instructional quality (Polikoff & Porter, 2014).

Teacher Expertise

The understanding of teaching as a complex task has lead to a refocus on the teacher as the expert in the "art of teaching," a perspective that differs from the process-product paradigm, which tended to identify isolated skills (Bromme, 2001). The teacher expertise paradigm in contrast is mainly interested in the cognitions of teachers and the way they processes information effectively (e.g. student reactions) and successfully responds to them. Although the teacher is the key person in expertise research, the goal is not to identify general personality traits of an effective teacher but to understand how teachers juggle the potentially taxing amount of information and make quick executive decisions regarding conflicting demands in their everyday teaching practice. As experts of teaching, the professionalism of teachers is emphasized, which opens the door to new empirical approaches of a research domain that usually investigates professionals whose expertise is not in question (pilots, musicians, athletes, chess players, etc.) and who are compared to novices who have at best rudimentary knowledge of the profession. In a study comparing biology experts and novices, for example, Jarodzka, Scheiter, Gerjets, & Van Gog (2010) were able to demonstrate using eye-tracking technology that experts were focusing more efficiently on presented video material to solve a categorization task—a finding in line with results of a recent meta analysis on expert–novice differences in eye movements (Gegenfurtner, Lehtinen & Säljö, 2011). Expert and novices are usually operationally defined based on experience or formal qualifications. For teacher expertise, researchers often compared experienced teachers with student teachers or

identify expert teachers through peer nomination (Bromme, 2001). The advantage of using external or formal criteria to identify experts that it avoids the difficult task to define what "good teaching" is. Teachers in training or student teachers at the beginning of their career make a natural comparison group even if the assumption holds that teachers longer on the job accumulate expertise (Borko & Livingston, 1989).

Teacher expertise as a research approach has helped with a problem in classic classroom environment research, namely, the fact that results substantially depend on the perspective (Fraser, 1991): Teacher self-rating, observer rating, and student ratings rarely converge in their assessment of the classroom when given parallel versions of a measure. Even after accounting for measurement error, the correlations rarely exceed moderate levels and are often close to zero, depending on the aspect under consideration (Ben-Chaim & Zoller, 2001; Kunter & Baumert, 2006). Applying research methods from traditional expertise research (Ericsson, 2006), an expert teacher is not expected to be able to reflect on her level of expertise and a novice teacher may or may not know in what ways he does not teach effectively. Teachers remain the key source of information but do not make self-assessment regarding teaching quality. Similarly, students are not expected to be capable of judging the quality of the classroom but are an important source of information regarding a learning criterion as proof of effective teaching, e.g. by providing test scores.

The expert–novice comparison research on teaching has revealed several characteristic differences that dovetail with Shulman's (1986) distinction of content knowledge (knowledge in the subject taught), pedagogical knowledge (general knowledge how to interact with students of a certain age), and pedagogical content knowledge (specific knowledge how to present content in a way that enhances understanding). Compared to novices, expert teachers have better cognitive representation of the content knowledge; they have extensive and flexible pedagogical content knowledge and can adjust their teaching more flexibly in response to changing needs of diverse learners. Experts are better at discerning relevant phenomena during a lesson. Based on prior experience they know, for example, to what extent student misbehavior can be tolerated or needs reprimand to avoid loss of academic focus. Expert teachers have established simple but effective routines early in the school year. Those routines are particular effective for standard activities like homework review, etc. (Berliner, 2001; Blömeke, Felbrich, Müller, Kaiser & Lehmann, 2008; Bromme & Dobslaw, 2003).

While the expertise paradigm has stimulated research on teaching by opening new avenues of investigations, it differs from research on expertise in other professions. Compared to teaching, it is easier to operationalize the concepts of expertise in other professions: what instrument a pilot focuses on in a critical situation and what action an expert takes is directly linked to potentially fatal outcomes in a flight simulator; what tissue a heart surgeon looks at closely and where and how she cuts compromised vessels directly affects the outcome of the operation. For classroom research, the distinction between professional and not-so-professional behavior is not always easy to make. What variables to focus on in a regular interactive classroom setting is less clear, and the positive effects of an expert teaching style in a math class compared to a novice teaching style is usually not immediately obvious or measurable. While traditional expertise research relies on low-inference "objective" measures like eye movements as indictor of cognitive focus (van Gog, Kester, Nievelstein, Giesbers & Paas,

2009), most research on teacher expertise is high inference or "subjective" in the sense that findings rely on the interpretation of traditional observational data, video footage, and/or teacher interviews.

Instructional Quality

A different empirical approach in classroom research is based on systematic classroom observations from a multidimensional perspective of classroom management. The tenet is that every classroom is characterized by a certain profile with respect to a limited list of learning-relevant aspects (dimension) of a regular classroom. The success or efficacy of a teacher is primarily measured using student outcomes like subject-specific tests or measures of criteria-based learning. The research then focuses on identifying those dimensions that are associated with higher rates of student learning. This approach holds promise to identify those aspect of teaching practice that are relevant for learning and strategies to improve those practices early in the careers of teachers. The use of observational tools for research purposes was further stimulated by the advent of video technology that made it possible to investigate quality aspects of instruction more objectively (Miller & Zhou, 2007). Many observational inventories were developed based on Flanders (1970), Interaction Analysis Categories (FIAC) or Brophy & Good's (1969), Teacher Child Dyadic Interaction System (TCDIS). Derivatives are available for specific school subject and various grade levels. The popularity of many of those measures is surprising given the weak theoretical foundation and lack of empirical validation for most of them, notably with the exception of scales for mathematics instruction (Kunter & Baumert, 2006; Hill, Rowan & Ball, 2005). One exception is the Classroom Assessment Scoring System (CLASS) observational tool that was developed by Robert Pianta and his colleagues over the course of more than two decades based on theoretical research and many empirical studies (Pianta, La Paro & Hamre, 2008). CLASS assesses classroom settings on ten components measuring emotional support, classroom organization, and instructional support. The validity of the CLASS coding system to predict validity for social and academic outcomes has been shown repeatedly (e.g. Allen, Pianta, Gregory, Mikami & Lun 2011; Curby, Rimm-Kaufmann & Abry, 2013; Hafen, Allen, Mikami, Gregory, Hamre & Pianta et al., 2012). Its objectivity and reliability is sustained by a mandatory training of coders. The CLASS coders do not measure expertise of teachers. They are certified experts in rating observed instruction on the CLASS dimensions, which are—as such—descriptive in nature. The advantage over the teacher expertise paradigm lies in the fact that no a priori criterion of teacher expertise is necessary and student teachers may score higher on CLASS dimensions than teachers with 20 years of teaching experience. A limiting factor for CLASS assessment is that it is usually not feasible to code more than a small sample of class periods, which have to be assumed representative of the classroom practice throughout the school year if CLASS codes are used as indicators of teaching quality and predictors of students' learning gains.

Like most observation tools, CLASS is a high-inference coding system, i.e. even with strong interrater reliability coding relies on the observer's interpretation of the student–teacher and student–student interactions as observed in the classroom or based on video footage. The advantage of reliable human coding, however, also makes it time

consuming and expensive. Hence, it is a tool that lends itself more to teacher evaluation or research on teaching, but it is less suitable for teacher training or professional development. Indicators based on information that is easy to collect and that can be analyzed more quickly would be particularly helpful in the context of professional development if there were empirical evidence of their validity (for an example, see Wang, Miller, & Cortina, 2013). The purpose of the current study was to investigate, based on an expert/novice design, whether teacher eye movements as a low-inference measure are related to CLASS dimensions as high-inference measures of instructional practice.

Low-inference measures are usually derived from physical measures, like sound, visual data, or movement patterns. The appeal is that those indicators are easier to measure "objectively," i.e. independent of coder training. An example of a relatively simple low-inference measure of classroom practice would be the time a teacher and the students talk throughout a class period. A simple recording can be used to measure both reliably. With the underlying assumption that teachers should try to limit their talking to make room for more elaborate students' responses, questions, and student-on-student discussions, this might be a simple but effective measure of teaching quality (Wang et al., 2013)—an idea that can be traced back to Flanders (1970).

The Current Study

With the current study, we want to demonstrate that teacher eye tracking as a low-inference measure can be used as valid indicator of teacher expertise. Different from teacher expertise research in the past, which was primarily based on high-inference qualitative approaches, our study brings research on teacher expertise closer to the tradition of professional expertise research that mainly uses low-inference measures to compare the cognitions and behaviors of experts and novices. Introducing low-inference measures into the field of teacher expertise research is not an end in and of itself. Low-inference measures have the potential to be developed into indicators of improvements in teaching that are easy to collect and inexpensive to analyze. This would allow for timely feedback and repeated measurement, making it a particularly helpful tool in the context of teacher training and professional development.

For the purpose of the current study, we used Mobile Eye-Tracking (MET) technology. In the past, eye-tracking research required an experimental lab setting where the subject was attached to a fixed apparatus that was able to record eye movements and to link it to the physical environment in front of the (Duchowski, 2007). Screen-based eye tracking where eye movement can be tracked while they watch video footage (e.g. Tobii Technology, Stockholm, Sweden) was an important improvement that opened possibilities to use eye tracking for research on teaching. Yamamoto & Imai-Matsumura (2012), for example, demonstrated using a Tobii video eye tracker that teachers who later correctly identified misbehaving students had focused on the target students more frequently and for longer durations when watching a 1-min video of first grade classroom. Using a similar video setup, van den Bogert et al. (2014) demonstrated that, compared to novices, experienced teachers distribute their attention more evenly across when watching video footage of a classroom situation.

With computer technology fast evolving, eye tracking has gone mobile with goggle-designed eyewear that allows educational researchers to use this technology in the field,

i.e. in regular classroom with minimal disruption of the instructional routine. The technology produces a digital video that shows the visual field of the teacher and a little marker indicating the current point of the teacher's visual focus. In line with traditional expertise research in other fields, the basic assumption is here that this low-inference measure is a valid indicator of teachers' attention focus.

There are various ways to use MET data for teacher training or professional development, starting from simply replaying it to the teacher to sophisticated technique of comparing eye movements of experienced teachers with those of novices. As it is true for any more rigorous use of video footage, it is necessary to reduce the complexity of the data. For the teacher MET data, we took advantage of the knowledge about the cognitive processing of visual information and limited the analysis on those instances when a teacher fixated on a certain point in the environment. The continuous video stream was reduced to a more manageable sequence of 3000–6000 fixations per 45-min lesson as basis of further analysis. The research question for the current study was: Does it really matter where a teacher looks during the lesson? Does the fixation pattern of a teacher reflect his or her level of teaching expertise? For the purpose of our study, we chose to validate an indicator derived from MET data with CLASS codes as an established high inference inventory of classroom quality. From a preliminary novice–expert teacher comparison, we knew that, compared to novices, expert teachers tend to focus their attention (a) more often on the students and (b) distribute their fixations on students more evenly across all students in the classroom (Miller & Correa, 2010; van den Bogert et al., 2014). The goal of our study was to investigate whether the distribution of the fixation among the students was associated with dimensions of the high inference CLASS dimensions of instructional quality. We hypothesized that a more even distribution of fixations on students is positively related to dimension of classroom organization, in particular behavior management. This expectation was based on the observation that student teachers tend to face more behavioral management challenges that may be associated with less effective monitoring of the classroom due to (over-) focusing on some students.

Method

Sample

The study is based on a subsample of a larger MET study that included 52 classroom teachers from 26 schools who volunteered to wear the MET gear during a regular class period. The schools were located in southeast Michigan and covered affluent neighborhood schools as well as schools in economically challenged neighborhoods. Two stationary cameras provided additional video footage. Half the participants were experienced teachers who mentored student teachers; half were their mentees, i.e. teachers who recently finished their university teacher training and taught the same class as their corresponding experienced teacher. Recordings for both teachers were made on different days with only one of the two teachers teaching. Classes varied in grade level (1–11) and school subject (Math, Science, English, and Social Studies).

For the current study, we selected 12 teacher pairs (12 experienced and 12 student teachers) from ten different elementary and middle schools from grade 2 to 8 with

steady seating arrangements throughout the lesson. First-grade classrooms were not included because they tend to vary seating within one class periods, which made the analysis of the eye-tracking material difficult because it is based on the coding of teacher fixations on individual students. High-school classrooms were excluded because the CLASS-coding rubric was not developed for the high school years, and the coders were not certified for those grades. Eighteen of the 24 teachers taught mathematics; the remaining 6 taught English.

Mobile Eye Tracking

Teachers wore the ASL Mobile Eye Tracker, a completely self-contained eye-tracking system (http://www.asleyetracking.com) that, in real time, integrates data of the location of the teacher's right pupil (using infrared recording) and the current visual field. The system consists of a pair of glasses containing infrared recording device and a small digital camera. A cable transmits the data to a recording unit that the teacher wears in a fanny pack around the hips. Two minutes prior to the lesson, we performed a 5-point system calibration in 5–10-m distance.

We first reduced the constant stream of eye-tracking video into a sequence of fixations with the assumption that fixation is a necessary condition for cognitive processing of visual information. Based on piloting, we defined a fixation as occurring when the gaze moved less than a defined distance (square root of 14 pixels) in either horizontal or vertical dimensions for a minimum of three samples (99 ms). A 45-min class period is reduced to approximately 4000 fixations. The number of fixations did not differ between expert and novice teachers. Using the time stamp for each fixation, trained assistants coded what the teacher was focusing on at the time of fixation. Coders used a list of eight standard codes (e.g. black board, instructional material, student material, etc.). When the teacher looked at a student, the student number was coded based on a classroom picture which arbitrary student numbers as reference. For the purpose of the current study, the important codes relate to the 15–32 students in the classroom. For each student, the fixation count was calculated. On average, a teacher looked at each student 115 times during a 45-min class period. However, these numbers vary substantially from student to student.

Data Processing

GINI Coefficient. The number of fixations per students in a 45-min class period varied substantially within and across classrooms. To capture the attention distribution, several measured could be used (range, variance, etc.). We used the GINI coefficient, which is frequently used in sociological and economic research as a measure of how uneven scarce goods are distributed in a population. The GINI coefficient (GC) is a more appropriate measure of inequality of distribution in this case because the measures for students are not statistically independent. If student A garners a lot of attention from the teacher, less is left to distribute among the others. The GC ranges from 0 (meaning in this case: all students have the same number of fixations) to 1 (= one students get all student fixations, all other students are ignored with zero fixations). While one could argue that a good teacher should monitor the classroom and hence distribute his or her attention across students, effective monitoring does not imply that a GC close to 0 is the

ideal of effective teaching because students will differ in their need for teacher attention. A good teacher will most likely tend more to those students who struggle with understanding the material. However, as an indicator of teaching quality, we maintain that, everything else being equal, a low GC in general indicates a teacher's competence in monitoring the classroom (Sabers, Cushings & Berliner, 1991; Sadler, 2006). For the expert–novice comparison, student needs are held constant because they are teaching the same group of students.

CLASS Coding. Each of the 24 classrooms was rated using the Classroom Assessment Scoring System (CLASS K-3, Pianta et al., 2007, 2008), an observational assessment tool that codes 10 dimensions in three broader domains emotional support, classroom organization, and instructional support. Emotional support included four aspects related to the quality of classroom interactions and sensitivity to children's emotions and interests: *positive climate* (e.g. respectful interactions), *negative climate* (e.g. expressed negativity), *teacher sensitivity* (e.g. need awareness), and *regard for student perspectives* (e.g. emphasis on children's interests). Classroom organization consisted of three dimensions that reflect control of behaviors and facilitation of learning: *behavior management* (e.g. redirection of misbehavior), *productivity* (e.g. time management), and *instructional learning formats* (e.g. engagement of children). Instructional support was measured by three dimensions reflecting the promotion of critical thinking and language: *concept development* (e.g. complex thinking), *quality of teacher's feedback* (e.g. feedback facilitates learning), and *language modeling* (e.g. language stimulation).

We used the video footage of the stationary camera located at the back of the classroom focusing on the teacher and the blackboard/screen. A research assistant operated the camera and followed the teacher by turning the camera if the teacher moved around the classroom. CLASS coding is recommended for video segments no longer than 30 min. Therefore, we analyzed two 20-min segments for each video in order to have two equally long segments for a regular class period. For two shorter class periods, only one segment was available (46 segments total). Two CLASS-certified coders rated each segment independently. Each classroom video was hence coded 4 times on all ten dimensions resulting in a total of 920 ratings. There were three coders, which meant that one coder rated both segments of a video. The design was completely balanced, i.e. all coder pairings were realized. Assignment of video segment to coders was random. Two videos were initially coded by all three coders and discrepancies discussed. The interrater correlation across all dimensions was r=.79, and similar across dimensions. After averaging the two ratings per segment, the correlation for the assessment of both segments is r=.84 with considerable variation across dimensions: The stability over the class period is highest for the assessment of negative climate (r=.90), quality of teacher feedback (r=.85), and language modeling (r=.82), and lowest for the assessment of regard for student perspective (r=.33), behavioral management (r=.41), and concept development (r=.49).

Additional Control Variables. Apart from teacher expertise (0=novice, 1=experienced), we included teacher gender and subject taught (1=Mathematics, 2=English) since we expected the eye-tracking patterns to differ by subject. Prior analysis of the data has not revealed gender differences, but we wanted to rule out an interaction with subject taught.

Results

The Gini coefficient (GC) was, as expected, lower in experienced teachers with an effect size of d=.76 (Table 1). The GC was not significantly different between mathematics and English classrooms. There was also no significant interaction. Teacher sex was not a significant factor (no main effect, no interactions). Across the ten CLASS dimensions, experienced teachers got significantly higher ratings in behavior management. As a trend, they also have a slightly higher negative climate rating. The GC is not significantly correlated for expert and novice teachers of the same class. Of the CLASS dimensions, only Language Modeling is correlated across expert and novice teachers (r=.68, p=.015). The latter is due to the fact that mathematics classes have lower Language Modeling Scores than English classes.

As Fig. 1 shows, the overall assessment profile of the observed classrooms is similar to the profile reported in the literature: Classrooms are characterized by a relative positive climate, average to positive classroom organization, and relatively weak instructional support (Casabianca, McCaffrey, Gitomer, Bell, Hamre & Pianta, 2013). The ratings for seven out of ten scales are significantly higher for the first segments than the ratings for the second segments (upper line) in particular with respect to the classroom organization subscales (behavioral management, productivity instructional learning format). That the quality rating is weaker in the second half of the class period is not related to level of expertise (expert/novice), sex of teacher, or subject (Math/English). The decline in observed quality of a lesson is not unexpected because new material is introduced usually in the beginning of a class period, while seat work is more common in the second half of a period (Doyle, 1983). Students' diminishing attention and behavioral issues tend to occur more frequently towards the end of a lesson resulting in lower quality scores, particularly in behavioral management and productivity.

Table 1 Comparing expert and novice teachers

Variable	*M* Expert	*SE* Expert	*M* Novice	*SE* Novice	$T_{(paired)}$ (df=11)	*p*	Effect size (*d*)	$r_{(E/N)}$	*p*
Gini coefficient	0.269	0.02	0.340	0.03	−2.07	0.032	−.60	0.10	0.767
Positive climate	5.083	0.32	5.264	0.26	−0.50	0.313	−.14	0.25	0.438
Negative climate	1.472	0.19	1.208	0.08	1.63	0.066	.44	0.49	0.104
Teacher sensitivity	4.701	0.21	4.375	0.25	0.95	0.182	.27	−0.11	0.736
Regard student perspective	3.840	0.33	3.847	0.26	−0.02	0.493	.00	0.30	0.341
Behavior management	5.576	0.17	5.011	0.25	2.48	0.016	.48	0.42	0.173
Productivity	5.257	0.18	5.194	0.23	0.22	0.415	.06	0.09	0.773
Learning formats	4.361	0.22	4.292	0.29	0.22	0.416	.06	0.27	0.405
Concept development	2.583	0.27	2.694	0.22	−0.36	0.363	−.10	0.21	0.519
Quality of feedback	2.931	0.28	2.972	0.21	−0.12	0.452	.04	0.09	0.785
Language modeling	2.764	0.21	2.972	0.15	−1.32	0.107	−.38	0.68	0.015

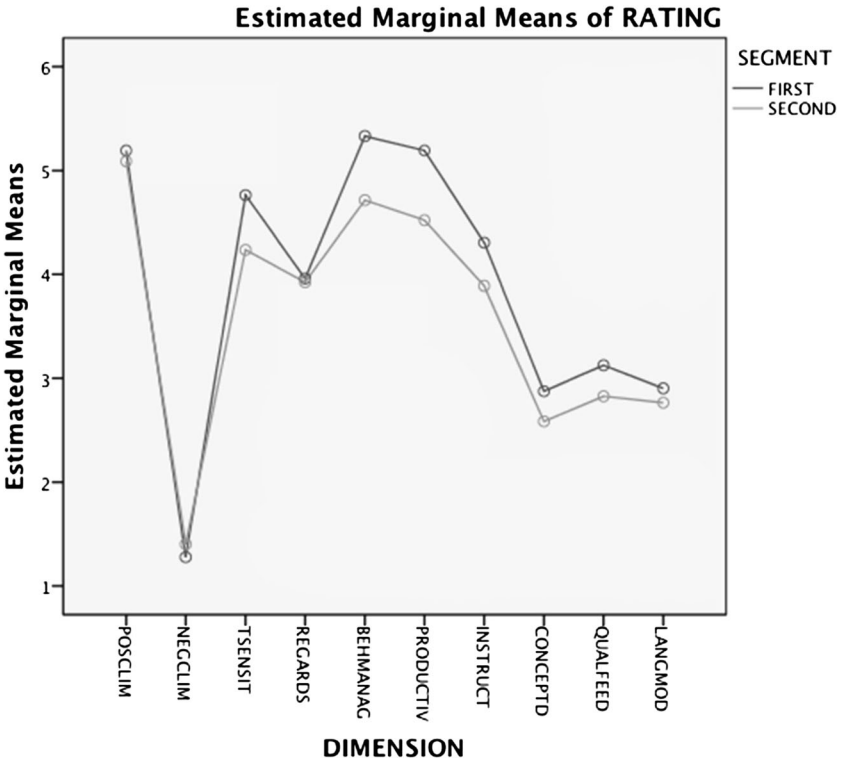

Fig. 1 CLASS rating by dimension and segment

Gini Coefficient and CLASS Dimensions

The primary research interest for this study was the link to the MET-derived Gini coefficient with the independent assessment along the CLASS dimensions. Table 2 shows that our primary hypothesis was not confirmed: Not only was there no significant association with GC and the three subscales of classroom organization, the

Table 2 Correlations CLASS dimension with Gini coefficient (GC); significant effect in italics

	GC	*N*	*p*
Positive climate	0.05	24	0.829
Negative climate	−0.26	24	0.221
Teacher sensitivity	0.09	24	0.685
Regard of student perspective	0.26	24	0.215
Behavior management	0.16	24	0.447
Productivity	0.33	24	0.117
Learning formats	−0.11	24	0.617
Concept development	0.22	24	0.294
Quality of feedback	*0.46*	*24*	*0.022*
Language modeling	0.23	24	0.276

correlations with two of them (behavioral management and productivity) were positive as a trend meaning that teachers got a higher productivity rating from the CLASS experts if they distribute the attention less evenly across the students in the classroom.

The only significant—again positive—correlation was with quality of feedback; teachers who provide more elaborate and useful feedback to students tend to have a higher GINI coefficient. Teachers who, with their feedback, engage students in further reflection on their responses, engage them in a short dialogue that helps expanding students' understanding are rated high in this category. A low score indicates that the teacher tends to simply dismiss incorrect answers and discourages students to share their reasoning for their answers. The unexpected positive correlation of feedback quality and GC is noteworthy not only because of its absolute size of r=.46, but it also turned out to be very robust: It remains within the limits of .23–.56 if the data were analyzed separately for the two segments or under three exclusion conditions (data of one of three CLASS raters at a time).

In an exploratory step, we reviewed all classroom videos that were in the top quartile in the GC (stronger focus on a small group of students in the class) and in the top quartile in the CLASS rating for feedback quality. The five identified videos were all taken with novice teachers, suggesting that the two substantive and significant effects regarding the GC, expertise, and CLASS quality of feedback rating are not independent. We created an interaction term following Aiken & West (1991) and introduced it as additional predictor in a multiple regression with GC as dependent variable.

All three effects were significant (Table 3).

Figure 2 depicts the three effects as a line graph. As expected the association between high GC and quality of teachers feedback exists exclusively for the novice teachers while experienced teachers have a low GC irrespective of their CLASS rating regarding feedback quality. Expert teachers are able to continue to monitor the classroom consistently even if they engage with one particular student to discuss his or her response.

Discussion

To our knowledge, this is the first empirical study that links a low-inference eye tracking indicator with high-inference assessments using CLASS, an established observational classroom assessment tool. Mobile eye tracking has only recently found its

Table 3 Interaction analysis

Source	Mean Square	*df*	*F*	*p* value
Intercept	2.18	1	424.13	.000
Expert	.031	1	5.95	.024
Quality of feedback	.038	1	7.37	.013
Expert × quality of feedback	.027	1	5.32	.032
Error	.005	20		

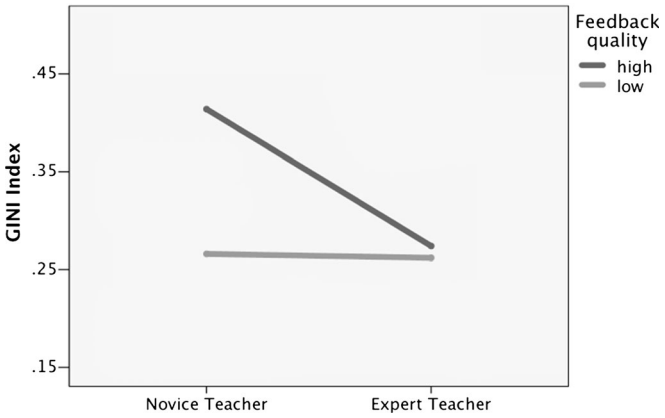

Fig. 2 Interaction expertise by quality of teacher feedback

way into classroom research due to the substantial progress in digital data processing that allows for a non-obstructive recording of teachers' eye movements in a regular classroom setting. Lacking prior studies to build on we started with the well-established observation that novice teachers find it more difficult than experienced teachers to manage the classroom with respect to students' time-on-task and distractive behavior. We hypothesized that one reason for this phenomenon is the tendency of novice teachers to focus their attention—measured as fixations on students—too much on a limited number of students, which invites other students to go off-task. While a plausible mechanism and supported by prior research (van den Bogert et al., 2014), the data did not confirm the underlying idea: The expected negative correlations of GC with desirable aspect of the instruction as reflected in CLASS codes were either not found or small in size.

The robust unexpected finding that quality of teacher feedback, i.e. teachers effort to engage students in elaborating on their responses, was associated with a higher GC does make sense in combination with the significant interaction with teacher expertise: Giving informative feedback to a student requires the teacher to give a particular student his or her undivided attention, which results inevitably in a higher number of fixations on that particular student—at the expense of fixations on classmates. This is reflected in the main effect Feedback Quality in the final regression analysis with GC as dependent variable. However, the interaction with teacher expertise suggests that experienced teachers are capable to still monitor the classroom with their eyes while giving individual students the additional attention when providing feedback. Revisiting videos where lots of feedback corroborated this impression: Student teachers tend to give feedback in a more "intimate" setting, i.e. walking close to the student, going on eye level with the student and lower their voice, while experienced teachers tend to give informative feedback openly with clear intention to share it with the entire class.

The last observation points to a dilemma with the attempt to combine low-inference ratings based on eye-tracking data and high-inference observational assessments: The latter is an overall rating that does not refer to specific teacher–student or student–student interactions (although CLASS coding protocol encourages raters to mention specific events). In fact, CLASS assessments can be strongly influenced by a single interaction that happened in a segment. Low-inference indicators, on the other hand,

refer to generalized patterns of attention throughout the entire segment or class period. While it would be possible to analyze situation-specific fixation counts, we refrained from doing so because the main idea of using MET as a tool for classroom research and teacher training is to develop a simple measure of general teaching practice and not a situation-specific measurement of behavior. The strength of MET lies in its face validity of what it measures: A novice teacher, for example, can review his/her eye tracking video to judge whether it would have been possible in a given situation to still monitor the rest of the classroom—maybe a student needed the teacher's attention but not to the extent given. This particular monitoring skill might be less obvious to an observer in the classroom or a coder watching video footage, but it can be identified easily using MET.

Our interpretation is in several respects speculative due to the limitations of our approach apart from the relatively small sample size. For example, the available data did not allow us to link teacher attention to student data, particularly student achievement, motivation, and social behavior. It would mean an important step forward in exploring the usefulness of MET data if attention could be analyze in relation to students' needs. The current study is somewhat simplistic in the assumption that an even distribution of attention is a sign of teachers' professionalism. This seems to be justified if the students in the class are homogenous with respect to their skill level. But in many of the classrooms recorded for this study, heterogeneity of the student regarding prior knowledge and behavioral instability was obvious in the video footage, and a strictly equal distribution of attention would be hard to justify in many classrooms. The idea that this indicator does, however, reflect at least to a certain degree a desirable skill of a teacher is confirmed by the results: While teaching the same body of students, the GC was significantly and substantially (Cohen's d=0.6) lower for experienced teachers compared to their novice counterparts.

As a technique, MET is not only a promising tool for classroom research but also as a feedback tool in teacher professional development. The "objectivity" of low-inference data is probably particularly appealing to student teachers: When they reviewed their own MET recording as part of the larger project, it always triggered stimulating discussions about their teaching practice. Our research underscores, however, the need to further explore the ultimate meaning of teachers' eye movements for the classroom interactions and the classroom atmosphere. Our findings using CLASS codings suggest a trade-off between improved feedback quality and classroom monitoring, but only for novice teachers. If this finding were replicated in other studies, the next step would be to develop feedback techniques that are considered of high quality but that do not absorb the teacher's attention more than necessary. Needless to say, MET would be the tool of choice to document the efficacy of those techniques.

References

Aiken, L. S. & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y. & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333*, 1034–1037.

Ben-Chaim, D. & Zoller, U. (2001). Self-perception versus students' perception of teachers' personal style in college science and mathematics courses. *Research in Science Education, 31*(3), 437–454. doi:10.1023/A:1013172329170.

Berliner, D. C. (2001). Learning about and learning from expert teachers. *International Journal of Educational Research, 35*, 463–482.

Blömeke, S., Felbrich, A., Müller, C., Kaiser, G. & Lehmann, R. (2008). Effectiveness of teacher education. *ZDM Mathematics Education, 40*, 719–734.

Borko, H. & Livingston, C. (1989). Cognition and improvisation: Differences in mathematics instruction by expert and novice teachers. *American Educational Research Journal, 26*, 473–498.

Brophy, J. E & Good, T. L. (1969). *Teacher–child dyadic interaction: A manual for coding classroom behavior. Report Series No. 27*. Retrieved from ERIC database. (ED042688)

Brophy, J. E. & Good, T. L. (1974). *Teacher-student relationships: Causes and consequences*. New York, NY: Holt, Rinehart and Winston.

Brophy, J. E. & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 328–375). New York, NY: MacMillan.

Bromme, R. (2001). Teacher expertise. In N. J. Smelser, P. B. Baltes & F. E. Weinert (Eds.), *International encyclopedia of the behavioral sciences: Education* (pp. 15459–15465). London, England: Pergamon.

Bromme, R. & Dobslaw, G. (2003). Teachers' instructional quality and their explanation of students' understanding. In M. Kompf & P. Denicolo (Eds.), *Teacher thinking twenty years on: Revisiting persisting problems and advances in education* (pp. 25–36). Liss, NL: Swets & Zeitlinger.

Campbell, R. J., Kyriakides, L., Muijs, R. D. & Robinson, W. (2003). Differential teacher effectiveness: Towards a model for research and teacher appraisal. *Oxford Review of Education, 29*, 347–362.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K. & Pianta, R. B. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*(5), 757–783. doi:10.1177/0013164413486987.

Curby, T. W., Rimm-Kaufman, S. E. & Abry, T. (2013). Do emotional support and classroom organization earlier in the year set the stage of higher quality instruction? *Journal of School Psychology, 51*, 557–569.

Doyle, W. (1983). Academic work. *Review of Educational Research, 53*, 159–199.

Duchowski, A. T. (2007). *Eye tracking methodology: Theory and practice*. London, England: Springer.

Ericsson, K. A. (2006). *Development of professional expertise: Toward measurement of expert performance and design of optimal learning environments*. Cambridge, UK: Cambridge University Press.

Flanders, N. A. (1970). *Analyzing teaching behavior*. Reading, MA: Addison-Wesley.

Fraser, B. J. (1991). Two decades of classroom environment research. In B. J. Fraser & H. J. Walberg (Eds.), *Educational environments: Evaluation, antecedents and consequences* (pp. 3–27). Elmsford, NY: Pergamon Press.

Gegenfurtner, A., Lehtinen, E. & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review, 23*, 523–552.

Hafen, C. A., Allen, J. P., Mikami, A. Y., Gregory, A., Hamre, B. & Pianta, R. C. (2012). The pivotal role of adolescent autonomy in secondary school classrooms. *Journal of Youth and Adolescence, 41*, 245–255.

Hill, H. C., Rowan, B. & Ball, D. B. (2005). Effect of teachers' mathematical knowledge for teaching on student achievement. *Americal Educational Research Journal, 42,* 371-406. Jarodzka, H., Scheiter, K., Gerjets, P., & Van Gog, T. (2010). In the eyes of the beholder: How experts and novices interpret dynamic stimuli. *Learning and Instruction, 20,* 146–154.

Kunter, M. & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environments Research, 9*, 231–251.

Miller, K. F. & Zhou, X. (2007). Learning from classroom video: What makes it compelling and what makes it hard. In R. Goldman, R. Pea, B. Barron & S. J. Derry (Eds.), *Video research in the learning sciences* (pp. 321–334). Mahwah, NJ: Lawrence Erlbaum.

Miller, K. F. & Correa, C. (2010, June). *Attention in the classroom. Teacher eye movements as an index of situation awareness.* Poster presented at the 5. IES research conference, National Harbor, MD.

Muijs, D. & Reynolds, D. (2001). *Effective teaching: Evidence and practice*. London, England: Sage.

Pianta, R. C., Hamre, B. K., Haynes, N. J., Mintz, S. L. & La Paro, K. M. (2007). *Classroom assessment scoring system manual, middle/secondary version*. Charlottesville, VA: University of Virginia.

Pianta, R. C., La Paro, K. M. & Hamre, B. K. (2008). *Classroom assessment scoring system, manual, K–3*. Baltimore, MD: Brookes.

Polikoff, M. S. & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis*. Advance online publication. doi:10.3102/0162373714531851.

Sabers, D. S., Cushing, K. S. & Berliner, D. C. (1991). Differences among teachers in a task characterized by simultaneity, multidimensionality, and immediacy. *American Educational Research Journal, 28*(1), 63–88.

Sadler, T. D. (2006). "I won't last three weeks": Preservice science teachers reflect on their student-teaching experiences. *Journal of Science Teacher Education, 17*(3), 217–241.

Seidel, T. & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77*, 454–499.

Shulman, L. S. (1986). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*, 4-14.

van Gog, T., Kester, L., Nievelstein, F., Giesbers, B. & Paas, F. (2009). Uncovering cognitive processes: Different techniques that can contribute to cognitive load research and instruction. *Computers in Human Behavior, 25*, 325–331.

van den Bogert, N., van Bruggen, J., Kostons, D. & Jochems, W. (2014). First steps into understanding teachers' visual perception of classroom events. *Teacher and Teacher Education, 37*, 208–216.

Wang, Z., Miller, K. F., & Cortina, K. S. (2013). Using the LENA in teacher training: promoting student involvement through automated feedback. *Unterrichtswissenschaft, 41*, 290-305.

Yamamoto, T. & Imai-Matsumura, K. (2012). Teachers' gaze and awareness of students' behavior: Using an eye tracker. *Innovative Teaching, 2*. Retrieved from http://www.amsciepub.com/doi/full/10.2466/01.IT.2.6.