First Results From a Validation Study of TAMI: Toolkit for Assessing Mathematics Instruction

Charles N. Hayward, Timothy Weston, and Sandra L. Laursen
University of Colorado Boulder

Many researchers consider observation to be a 'gold standard' for measuring classroom practices since self-report surveys may be prone to bias. In this paper, we explore how the design of survey instruments and observation protocols affects the trustworthiness of the data collected. We describe our process of developing well-aligned observation and survey instruments in order to reduce sources of measurement error. We present results from a large-scale test of these instruments in 176 observations of 17 different math courses. Our results indicate that when survey instruments are designed to describe what happens in a course, rather than evaluate the quality of the instruction, and when those survey results are compared to observation protocols measuring teaching in the same way, self-report surveys are largely trustworthy.

Keywords: Measurement, Instruction, Observation, Surveys

Describing and assessing instructional practices in science, technology, engineering, and mathematics (STEM) courses is a difficult undertaking. Various methods exist, each with their own advantages and disadvantages (American Association for the Advancement of Science, 2012). Two commonly used methods in RUME studies are classroom observation and instructor surveys. Observational data is collected by a neutral third party so it is sometimes considered more objective or 'accurate' than self-reported survey data, which may be prone to subjective bias (Ebert-May, Derting, Hodder, Momsen, Long, & Jardeleza, 2011).

However, observations have their own sources of measurement error. Observers often need significant amounts of training to ensure sufficient inter-rater reliability (IRR) – the degree to which different raters agree on their ratings of what they observe (Smith, Jones, Gilbert, & Wieman, 2013). Though IRR is usually used as a standard for judging quality of observation data through adherence to a particular protocol, IRR only captures whether observers apply the protocol consistently, and does not address other inherent sources of variability in observation data. Multiple observations of the same instructor are needed to make confident inferences about their teaching as a whole (Pianta & Hamre, 2009), since two or three class sessions may not represent the entire class (Hill, Charalambous, & Kraft, 2012). But if more observations are needed, this adds significantly to the time and cost of collecting observation data. This drives the need for validity comparisons (Hill, Charalambous, & Kraft, 2012): if researchers can confidently substitute a survey for observations, then they can increase the number of participants involved while decreasing the costs and time invested.

A number of studies have attempted to answer the question of whether survey data can be as trustworthy as observations, and they have come to different conclusions. Ebert-May et al. (2011) found that while most instructors said in surveys that they changed their courses to become more active and learner-centered, most were observed using traditional lectures and teacher-centered instruction. A similar study conducted in a K-12 context (Fung & Chow, 2002) found a mismatch between the teachers' conceptions of their teaching style and observed practices with teachers again overestimating the interactivity and student-centered characteristics of their teaching. However, college faculty members studied by Smith et. al (2014) were "generally self-aware" of how often they used methods related to lecturing and presentation.

If observations are considered more objective, mismatches are interpreted as 'inaccuracy' or 'bias' in self-report. However, it may be that mismatches are at least partially caused by a misalignment between survey and observation instruments. In this paper, we present results of our own validation study comparing an instructor survey and observation protocol. We first describe the framework and process we used to create two well-aligned instruments, and then present results from a large-scale study using them. We explore the questions:

1. What design choices affect alignment between surveys and observation protocols?
2. When using well-aligned instruments, in what ways does instructor self-report of instructional practices agree or disagree with observation data?

Conceptual Framework: Types of Observation Protocols

Observation protocols are characterized along multiple dimensions (Hora & Ferrare, 2012). The two main dimensions are descriptive vs. evaluative and segmented vs. holistic. Descriptive protocols aim to simply capture or describe what is happening in a class, whereas evaluative protocols rate the quality of a class. Segmented protocols divide a class into short time segments, usually 2 or 5 minutes long, with coders recording features of the class during each period. Thus, the whole class is characterized by the sum and sequence of short segments. In contrast, holistic protocols aim to characterize the class as a whole. The observer may take notes during class then use the notes as evidence to rate the class across multiple criteria. Additionally, protocols may focus on the instructor or students, may or may not take subject matter into account, may require high or low inference by the observer, and may vary in the degree of structure.

These dimensions can be combined in many different ways, but the two main dimensions capture the largest differences between protocols. A descriptive protocol may measure how frequently a practice such as group work occurs using a segmented approach (e.g. group work occurred during 17 of 25 two-minute intervals) or a holistic approach (e.g. "about ¾ of class time"). Evaluative protocols may measure the quality of group work by asking how students engaged in group work or whether the group task was structured effectively. Again, that can be done in a segmented way (e.g., marking student engagement during each interval) or holistically (e.g., rating on a 0-4 scale whether "students were productive and engaged in group activities").

	Segmented	**Holistic**
Descriptive	**TDOP** (Hora & Ferrare, 2013) **COPUS** (Smith, Jones, Gilbert, & Wieman, 2013) **RIOT** (West, Paul, Webb, & Potter, 2013)	**SPROUT** (Reimer, Schenke, Nguyen, O'Dowd, Domina, & Warschauer, 2016)
Evaluative	**PORTAAL** (Converse, Eddy, & Wenderoth, 2014)	**RTOP** (Sawada, et al., 2002) **MCOP2** (Gleason, Livers, & Zelkowski, 2017) **M-Scan** (Walkowiak, Berry, Meyer, Rimm-Kaufman, & Ottmar, 2014)

Figure 1. Framework for classifying observation protocols.

In Figure 1, we characterize some common observation protocols used in STEM courses based on their main design features; individual items may fall into the other quadrants. For example, some items on the MCOP2 are more descriptive than evaluative. Segmented protocols tend to be very detailed and granular, whereas holistic protocols zoom out to capture a broader view. The choice of a protocol should match the goals for its use. For example, if the goal is to offer formative assessment for instructor growth, a descriptive protocol such as RIOT, which focuses on instructor/student interactions, offers a lower-stakes measure conducive to constructive discussion. In contrast, RTOP is more evaluative and may be viewed as judgmental rather than constructive.

Surveys items about instructional practices can also be classified as descriptive or evaluative. Surveys may ask instructors to describe and quantify their behaviors over a period of time (e.g. "How often did you lecture this semester?"), or ask them to reflect on and evaluate their own teaching based on criteria such as the quality of their interactions with students, the types of activities used in the course, or their perceived ability to explain difficult concepts.

Methods

Now, we describe how we used this framework to design a descriptive instructor survey along with a well-aligned segmented, descriptive observation protocol. We then describe how we collected and compared data from both instruments in college math courses. The instruments are part of the Toolkit for Assessing Mathematics Instruction (TAMI) that we are developing.

Development of the Survey Instrument

Our survey came out of prior work evaluating professional development workshops (Hayward, Kogan, & Laursen, 2016). Our goal was to assess initial changes in the use of particular instructional practices, not to evaluate how well instructors were using these practices, since these skills may take years to develop. So our questions asked instructor to report the frequency of use of different practices. We designed our survey to be administered shortly after the conclusion of a course and to use descriptive rather than evaluative items, asking "*what did you do*?" versus "*how well did you do it?*" For this project, we conducted think-aloud interviews to adjust the items and answer choices to better align with instructors' conceptualizations of their own practices.

On our final survey, instructors first report how frequently they use 11 different classroom practices commonly seen in college math courses including group work, whole class discussions, formal lecture, interactive lecture, and student presentations. Frequencies are measured with a 7-point scale with concrete descriptors from 'never,' to 'about once a month,' to 'every class.' Then, instructors report the duration of use for each of the practices they used: 'a few minutes,' '1/4 class,' '1/2 class,' '3/4 class,' or 'entire class.' Open-ended items ask instructors to describe patterns in practices or rare events (e.g. computer lab for the last 3 classes.) Finally, instructors supply text descriptions of 'lecture,' 'presentations,' and 'group work' in the courses.

Development of Observation Protocol

Background. A 2011 paper by Ebert-May et al. is often used to argue that self-report is not 'accurate.' They compare biology instructors' self-reported practices to observations coded with the Reformed Teaching Observation Protocol (RTOP) (Sawada, et al., 2002). Many instructors reported using active learning strategies, but RTOP scores were more in line with lecture-based teaching. The authors concluded, as their title suggests, that 'what we say is not what we do.'

However, our own analysis of the RTOP instrument reveals that this protocol is not well aligned with the data collected through self-report. The RTOP is evaluative and holistic, whereas the self-reported survey items are more descriptive and behavior-oriented. We were curious how much of the discrepancy between observation and survey methods was due to a difference in what is being measured, rather than an 'inaccuracy' in self-report. In other words, were the answers really different, or were they comparing answers to two different questions?

Design of our observation protocol. Our existing survey functioned well for evaluation work, and internal consistency provided evidence that it was trustworthy (Hayward & Laursen, 2014). It is a descriptive, frequency-based behavioral survey, so we wanted a descriptive, segmented, behavior-oriented observation protocol to match. Measures in the TDOP (Hora & Ferrare, 2013) offered more detail than we needed, but the COPUS (Smith, Jones, Gilbert, & Wieman, 2013), a simpler modification of the TDOP intended for STEM courses, did not quite align with practices we saw in college mathematics courses, especially with its focus on clicker use.

We modified the COPUS by changing a few codes and adding others to better align it with practices we saw in mathematics classes and with survey items we previously developed through interviews with mathematics instructors. We incorporated the ICAP framework (Chi & Wylie, 2014) as a research-based measure of the nature of student engagement. We included some end-of-class holistic items that are evaluative and descriptive. Thus, while most of our protocol is segmented and descriptive, it also includes items from each of the other quadrants. The main portion aligns well with our survey items, but adding items from other quadrants allows us to analyze how misalignment may affect the comparison between survey and observational data.

Sample

Our sample included 176 in-person class observations from 17 courses. Our average of 10.4 sessions per course is many more than is typical (1-3 observations per course), but was necessary to ensure that we obtained a truly representative sample (Weston, Hayward, & Laursen, 2017). The data included 4789 two-minute observations, or nearly 160 hours of observations carried out over two terms at three public universities in courses on algebra, calculus, geometry, statistics, and mathematical modeling. Class meetings were 50 minutes long, meeting three or four times per week, or 75 minutes long, meeting twice per week. All courses were on semester schedules.

Reliability of Observations

In piloting our observational protocol we assessed inter-rater reliability (IRR). Overall IRR was high, with raters agreeing on 93% of their observations over each two-minute period and varying only modestly by the type of item. Modest variations were also found for how well raters agreed when rating activities for different teachers, from 91% to 96% depending on which teacher was observed. These results are on par with those of other published protocols.

Analysis Methods

Comparing survey to observational results was complicated by differences in the frequency of measurement, with surveys given once at the end of each term, and observations taking place multiple times throughout the term. To make a fair comparison, we aggregated observations at the course level by taking averages across all classes observed. We also aggregated within similar types or formats of classes such as classes primarily devoted to lecture, group work or a mix between these formats. Aggregating this way makes a fair comparison to survey items that ask instructors to estimate the proportion of time spent in classes "*when you used this activity.*"

We made two types of comparisons. First was the comparison of the instructors' report of average amount of time within each class devoted to specific activities (such as lecture) compared with the average observed time devoted to this activity. Observational averages took values between 0 and 100%; survey values asked teachers to estimate rougher proportions of class time spent doing the activity (e.g., "entire class," "¾ of class"). For this analysis we used a fairly liberal criterion and considered bivariate points a match if the observational value fell between the two nearest boundaries for survey proportions – these 'match' ranges are represented as vertical green bars in Figures 2-4. For instance, if an instructor estimated he lectured ½ of the class, any observed value between the next nearest survey responses of 25% and 75% was considered a match. Lower and higher boundaries were set at 10% and 90% of class time. Analysis with the full data set will use correlation coefficients and other tests of congruence such as the Kappa coefficient.

Second, we created an interactivity index based on the number of questions teachers and students asked during lecture. To create a three-point scale aligned with our three categories, "formal," "some interaction," and "interactive," we counted frequencies for six question/answer types – 2 for students and 4 for instructors. Each item was scored as its tertile (1,2, or 3) of the frequency distributions. We averaged these scores, and again split into tertiles for the final index.

Results

We compared observation and survey data from all participants who were observed in-person, totaling 176 observations of 17 courses, although most comparisons used 13 or 14 courses depending upon which activities were reported. Results from this analysis are presented below. We are currently integrating data from 141 additional observations from 16 courses observed via video camera and results from the full analysis will be available by February.

We first examined the match between survey responses and observational data for lectures. The survey question about lecture asked: "*On average, when you used this method, did you use it: Entire class, ¾ class, ½ class, ¼ class, a few minutes."* We asked about three types of lecture: formal (little or no question or answer), some interaction, and frequent interaction. We compared the dominant mode of lecture reported in surveys to the observed averages of time spent lecturing in classes (Figure 2). For the most part, instructor estimates were aligned with what we observed. Four out of 13 cases were considered misclassifications, for a "true" classification rate of 69%. However, two of these errors were very near the classification boundaries. One instructor drastically underestimated the amount of lecture used, relative to what was observed.

We also compared survey instructor ratings of their own lecture interactivity to observed interactivity in classes (Figure 3). The tertiles for the interactivity index were 1–2, 2–2.5, and 2.5-above. Again we saw moderately high congruence between how instructors rated the interactivity of their courses and observed interactivity. Three out of 14 teachers highly overestimated the interactivity in their classes, and one teacher slightly underestimated the interactivity of his/her teaching. Overall accuracy was 69%.

We also compared averages for instructors' reported use of group work (Figure 4). We found that most instructors were at the extremes – either they mixed a fairly small amount of group work with lecture or other activities, or devoted the whole class to working with groups, usually on one day of the week devoted to recitation. Two instructors slightly overestimated the amount of time their students spent working in groups, one slightly underestimated, and one made a large overestimation. Overall accuracy was 71%.

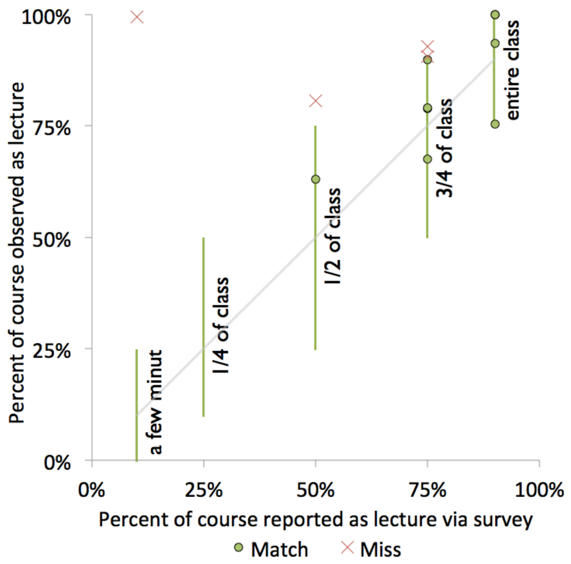

Figure 2. Comparison between survey and observations of lecture.

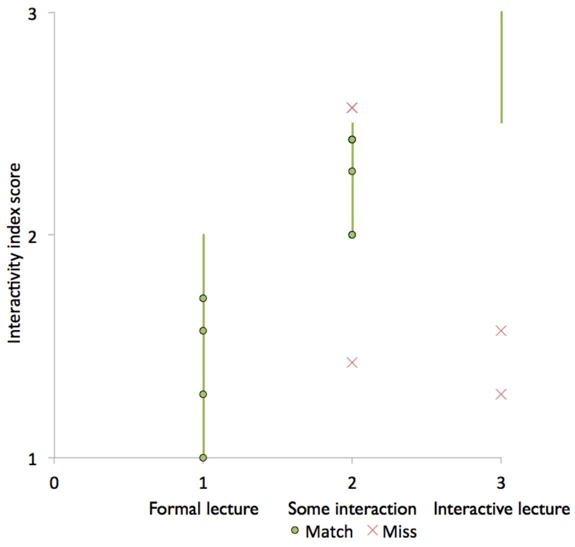

Figure 3. Dominant lecture mode defined by survey and interactivity index from observations.

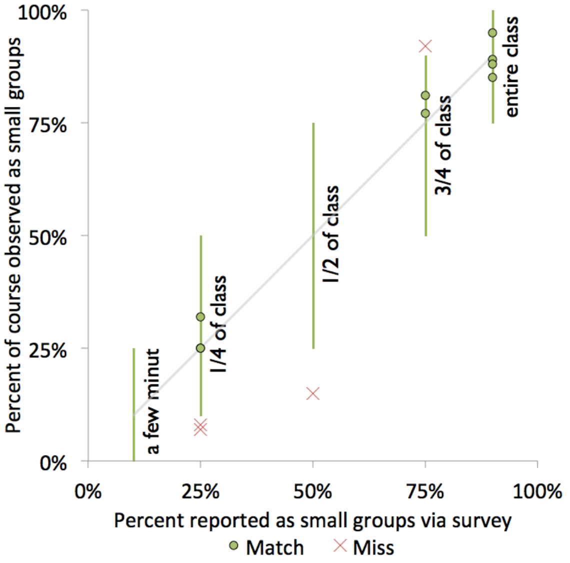

Figure 4. Comparisons between survey and observations of group work.

Discussion

We made three validity comparisons with our initial data. Overall, it seems that when survey and observational measures are aligned, instructors' self-reported practices are aligned with observation data. For the two comparisons of proportions of time spent lecturing and in structured group work (Figures 2 and 4), most instructors seem to have an accurate idea of the proportion of class time spent on each activity. These descriptive comparisons are fairly straightforward, and instructors likely remember their basic lesson structure over the course of a term. Not surprisingly, some underestimate the amount of class time they spend lecturing, and overestimate the time students work in groups. However, most of these differences are relatively small and are related to extreme values; at the upper ends, instructors tended to underestimate and at the lower ends, they tended to overestimate. Only one participant had a large discrepancy, which may be due to differences in how our definition of "lecture" differed from the instructor's.

Estimates of teacher-student interactivity (Figure 3) were also fairly accurate in terms of the number of matches. However, the discrepancies are large; those who reported the most interaction were some of the lowest-rated in observations. These results are interesting when interpreted through the design framework. Our interaction index is a norm-referenced measure. This means ratings are based on comparisons to other courses in the dataset rather than to an outside, objective standard. When we collapsed the three types of lecture on the survey (formal, some interaction, and interactive), self-reports aligned well with observations. However, when we used a more evaluative approach by comparing the type of lecture with our interaction index, discrepancies were large. So while instructor self-report was quite accurate with strictly descriptive measures (i.e. duration of lecture or group work), there were greater discrepancies with this more subjective, evaluative index. Past studies claiming that instructors were not 'accurate' in self-report relied heavily on evaluative measures, and our results suggest that using descriptive instruments instead of evaluative may help reduce these discrepancies.

Although observations are commonly thought to be objective, it is impossible to remove all forms of bias. We found that using segmented, descriptive items helps to reduce bias. However, protocol designers still must decide how to define items. Their perspectives bias what 'counts' for items. For example, when coding question and answers, we designed our protocol to only count when the instructor provided a real opportunity for students to respond. Many times, instructors asked rhetorical questions, or answered their own questions so quickly that students had no opportunity to respond. So we did not count these as questions. It is entirely possible that instructors frequently used these types of questions, but our coding would reflect very little interaction. The 'error' for those who drastically overestimated the interactivity of their lectures relative to our observations may originate in our definition of what counts as a question.

Despite multiple claims that self-report is not 'accurate,' the issue of trustworthiness is much more nuanced than how well it compares to observation data. Survey data may be prone to self-report bias, but there are also many sources of variation or error in observation data. These include observation protocol and survey design alignment, coding definitions of what 'counts,' and variability in day-to-day activities and representativeness of the observation sample compared to the whole course. Our results suggest that when survey and observation instruments are well designed and properly aligned, surveys may be a trustworthy, efficient, and less costly method of measuring teaching practices.

It remains an open question whether it is possible to use a survey to measure changes in teaching practice following professional development interventions. Issues of bias increase when instructors are expected to change their practices, and some may consciously or unconsciously overestimate the time spent on inquiry-based activities, and underestimate their time in teacher-centered lecture. Survey items must also be sufficiently sensitive to capture differences before and after the intervention. Future work should test the surveys in such intervention conditions.

Acknowledgments

This work was supported by a grant from the National Science Foundation (NSF DUE 1245436). Some data was collected and coded with support from the Spencer Foundation (*Taking the Measure of Math Class: Comparing Measures of Teaching in Undergraduate Mathematics Courses*, 2015). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Spencer Foundation.

References

American Association for the Advancement of Science. (2012). *Describing & measuring undergraduate STEM teaching practices: A report from a national meeting on the measurement of undergraduate science, technology, engineering and mathematics (STEM) teaching*. 2013: AAAS.

Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219-243.

Converse, M. S., Eddy, S. L., & Wenderoth, M. P. (2014, Nov 19). *PORTAAL Manual: Practical Observation Rubric to Assess Active Learning, An evidence-based classroom observation tool*. Retrieved August 10, 2017, from https://sites.google.com/site/uwbioedresgroup/research/portaal-resources

Ebert-May, D., Derting, T. L., Hodder, J., Momsen, J. L., Long, T. M., & Jardeleza, S. E. (2011). What we say is not what we do: Effective evaluation of faculty professional development. *BioScience*, 61(7), 550-558.

Fung, L., & Chow, L. P. (2002). Congruence of student teachers' pedagogical images and actual classroom practices. *Educational Research*, 22(3), 313-321.

Gleason, J., Livers, S., & Zelkowski, J. (2017). Mathematics Classroom Observation Protocol for Practices (MCOP2): A validation study. *Investigations in Mathematics Learning*, 9(3), 111-129.

Hayward, C. N., Kogan, M., & Laursen, S. L. (2016). Facilitating instructor adoption of inquiry-based learning in college mathematics. *International Journal of Research in Undergraduate Mathematics Education, 2*(1), 59-82.

Hayward, C. & Laursen, S. (2014). *Evaluating professional development workshops quickly and effectively.* 17th Annual Conference on Research in Undergraduate Mathematics Education. Denver, CO, February 27-March 1.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater relability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, 41(2), 56-64.

Hora, M. T., & Ferrare, J. J. (2012). *A review of classroom observation techniques used in postsecondary settings.* [White paper]. Prepared for the Measurement of Teaching Practices in Undergraduate STEM workshop hosted by AAAS/NSF.

Hora, M. T., & Ferrare, J. J. (2013). Instructional systems of practice: A multidimensional analysis of math and science undergraduate course planning and classroom teaching. *Journal of the Learning Sciences*, 22(2), 212-257.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher*, 38(2), 109-119.

Reimer, L. C., Schenke, K., Nguyen, T., O'Dowd, D. K., Domina, T., & Warschauer, M. (2016). Evaluating promising practices in undergraduate STEM lecture courses. *RSF: The Russell Sage Foundation Journal of the Social Sciences*, 2(1), 212-233.

Sawada, D., Piburn, M. D., Judson, E., Turley, J., Falconer, K., Benford, R., et al. (2002). Measuring reform practices in science and mathematics classrooms: The Reformed Teaching Observation Protocol. *School Science and Mathematics*, 102(6), 245-253.

Smith, M. K., Jones, F. H., Gilbert, S. L., & Wieman, C. E. (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.

Smith, M. K., Vinson, E. L., Smith, J. A., Lewin, J. D., & Stetzer, M. R. (2014). A campus-wide study of STEM courses: New perspectives on teaching practices and perceptions. *CBE-Life Sciences Education*, 13(4), 624-635.

Walkowiak, T. A., Berry, R. Q., Meyer, J. P., Rimm-Kaufman, S. E., & Ottmar, E. R. (2014). Introducing an observational measure of standards-based mathematics teaching practices: Evidence of validity and score reliability. *Educational Studies in Mathematics*, 85(1), 109-128.

West, E. A., Paul, C. A., Webb, D., & Potter, W. H. (2013). Variation of instructor-student interactions in an introductory interactive physics course. *Physics Review Special Topics - Physics Education Research*, 9(1), 010109.

Weston, T. J., Hayward, C. N., & Laursen, S. L. (2017). *When seeing is believing: Conditions for making confident inferences about teaching in semester-long courses from time-sampled classroom observations.* Manuscript in preparation.