Methodological development of a new coding scheme for an established assessment on measurement uncertainty in laboratory courses

Benjamin Pollard
Department of Physics, University of Colorado Boulder, Boulder, CO 80309, USA and
JILA, National Institute of Standards and Technology, Boulder, CO 80309, USA

Robert Hobbs
Department of Physics, Bellevue College, Bellevue, WA 98007, USA

Dimitri R. Dounas-Frazer
Department of Physics and Astronomy, Western Washington University, Bellingham, WA 98225, USA

H. J. Lewandowski
Department of Physics, University of Colorado Boulder, Boulder, CO 80309, USA and
JILA, National Institute of Standards and Technology, Boulder, CO 80309, USA

Student understanding around measurement uncertainty is an important learning outcome in physics lab courses across the US, including at the University of Coloroado Boulder (CU), where it is among the major learning outcomes for the large introductory stand-alone physics lab course. One research tool for studying student understanding around measurement uncertainty, which we use in this course, is the Physics Measurement Questionnaire (PMQ), an open-response assessment for measuring student understanding of measurement uncertainty. Interpreting and analyzing PMQ data involves coding students' written explanations to open-response questions. However, the preexisting scoring scheme for the PMQ does not fully capture the breadth and depth of reasoning contained in our students' responses. Therefore, we created a new coding scheme for the PMQ based on responses from our students. Here, we document our process to develop a new coding scheme for the PMQ, and describe the resulting codes. We also present examples of what can be learned from applying the new coding scheme at our institution.

I. INTRODUCTION AND BACKGROUND

Lab experiences are a key feature of most undergraduate physics programs. They allow students to engage with concepts and practices that are fundamental to physics, and STEM more generally. As such, education researchers have developed, and are currently developing, many assessment tools that are suited for lab courses, both for instructors to measure and improve their courses, and for researchers to understand the unique learning that occurs in lab settings [1–5].

One common learning goal in introductory physics labs concerns measurement uncertainty [3, 4]. Estimating uncertainties for measurements and using them when interpreting results is critical to the more general practice of modeling in experimental physics [6, 7], and thus situates measurement uncertainty as a fundamental aspect of experimental practice. As such, there have been several assessments used by physics education researchers that focus on students' facility with measurement uncertainty concepts and practices, for example [5, 8–11]. They range in scope and in their degree of focus on measurement uncertainty. We focus here on the Physics Measurement Questionnaire (PMQ) [8], an established assessment instrument used to study students' understanding of measurement uncertainty in physics lab courses.

The PMQ was developed over a decade ago in the context of a research project concerning lab curriculum reform at the University of Cape Town, Cape Town, ZA [8]. Starting from earlier work involving pupils aged 11-16 in York, UK [12], researchers in Cape Town found that the instruments developed in York were not suitable for their first-year university students, and thus developed a new set of survey questions for their own national and institutional context [8]. These questions, or probes, compose the PMQ survey itself.

Each probe concerns a particular aspect of measurement in the context of an experiment involving rolling a ball down a slope and then measuring the distance it travels in free-fall. These aspects include data collection, data processing, and data comparison. An example of one of these probes is shown in Fig. 1. It asks the student to make a decision involving how to represent a set of data. The data include two repeated values, hence the probe's name, UR, for "using repeats." As with all of the PMQ probes, the student makes a choice and then writes an explanation of their choice. Those two responses, the choice and the explanation, together form the data collected from each PMQ probe.

The Cape Town researchers initially interpreted and analyzed data from the PMQ using a theoretical model from the work in York [13]. However, they realized that the model lacked explanatory power and failed to capture the greater sophistication behind some of their students' responses [8], so they developed a new model based on data from their own context [14, 15]. Their model centers around two paradigms of measurement: the point paradigm and the set paradigm. The point paradigm holds that a single measurement can represent the true value of a physical quantity or measurand, and that values from individual measurements can be considered independently of each other. In contrast, the set paradigm

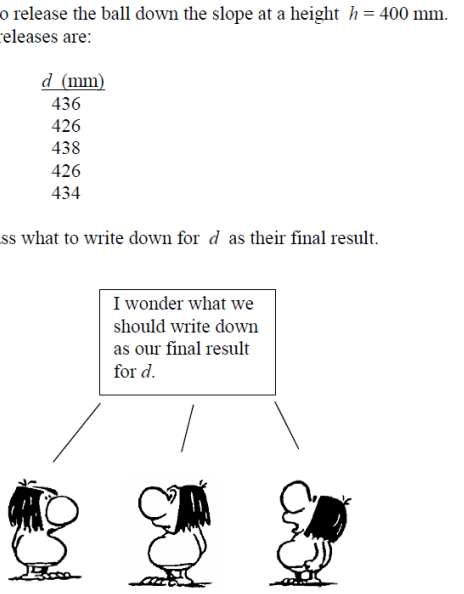

FIG. 1. The UR ("using repeats") probe of the PMQ. Students are prompted to "Write down what you think the students should record as their final result for d," and then to "Explain your choice."

recognizes that no individual measurement can yield the true value of the measurand, and that multiple measurements will form a distribution, providing increasingly more information about the measured quantity. The set paradigm tends to be more aligned with a probabilistic approach to measurement uncertainty [15], and characteristic of expert-like reasoning.

Our past work used the PMQ and the paradigm model to measure learning in an introductory physics lab course at the University of Colorado Boulder (CU) in Boulder, Colorado, US. Despite our context differing from that in which the PMQ was developed, we found that the PMQ nonetheless prompted responses from our students that could be analyzed using the point and set paradigms [16, 17].

At the same time, we also observed a greater variety and depth in our students' responses than could be captured by the paradigms, and realized that in order to better understand our students' reasoning, we needed to extend the paradigm model. To that end, we document here the creation of a new coding scheme for the PMQ, based on responses from students in the introductory physics lab course at CU, and present preliminary results from students in that course. We aim to achieve two goals: (i) to explicate how we developed a new coding scheme for the PMQ that is suited for use in a large introductory physics lab course at our US institution, and (ii) to demonstrate the deeper insight offered by the new coding scheme through preliminary results from that course.

II. CONTEXT AND METHODS

Our new coding scheme for the PMQ was developed in the context of a stand-alone large-enrollment lab course at CU, which students typically take during their second semester of study as their first physics lab course at CU. Most students have completed an introductory course in mechanics and are

concurrently taking an introductory course on electricity and magnetism. Self-reported information regarding gender, race, ethnicity, and major from the students who completed this course in Spring 2018, a semester studied here, can be found in ref. [18]. More information about this course can also be found in refs. [16, 17, 19].

As one of the instructors of the course, HJL administered the PMQ electronically at the beginning (pre) and end (post) of the course in every semester from Fall 2016 through Spring 2018. In all semesters, both pre and post surveys counted as assignments that were graded for participation only, representing 1-2% of the final course grade. The PMQ assignment was completed in class through Spring 2017. A transformed version of the course was taught starting in Fall 2018, and from that semester onward the PMQ was completed outside of class using a link provided via email. Our analysis of PMQ responses from this course, including the development and application of a new coding scheme for those responses, occurred over the same time period and continues to the present.

To develop a new coding scheme for the PMQ, RH initially coded anonymized responses collected in Fall 2016, with input from others on the research team. Only four of the PMQ probes were considered, and they remain the focus throughout this work. Those probes were RD (for "repeated distance"), UR (for "using repeats"), SMDS (for "same mean different spread"), and DMSS (for "different mean same spread"). We chose not to use the other PMQ probes because they were either incompatible with the electronic administration format, considered less useful by the researchers from Cape Town, or did not appear on both the pre-test and post-test versions of the PMQ. Starting with the original codes developed in Cape Town [20], RH added additional codes generated inductively to represent the unrepresented lines of reasoning in our data. After the set of codes was expanded, the research team together assigned a paradigm to each code: point, set, or neither/unknown if the responses represented by a code were insufficient to determine the students' underlying reasoning paradigm. Results from analysis of those assigned paradigms was reported in refs. [16, 17].

After that initial analysis, BP, HJL, and DRDF consolidated the set of expanded codes, grouping them thematically to create a scheme with fewer codes and more detailed descriptions. Doing so made the scheme more tractable as a research tool. As with the original coding scheme from Cape Town, we created a separate set of codes for each of the four PMQ probes we studied. Each code includes a two-character identifier, a name, and a definition. Each code is explicitly associated with a paradigm, represented by the first character of the identifier (P, S, or U). We allow multiple codes to be assigned to a single response.

This consolidated set of codes was first applied and refined using PMQ responses collected in Fall 2017. A random, anonymized subset of 20 responses to a single probe was collaboratively coded by RH and BP in order to elucidate ambiguities in the code definitions. After discussing and refining the code definitions together, RH and BP separately coded an additional 40 anonymized, random responses, ensuring that none of the first 20 were included in this second subset. We calculated the Cohen's kappa statistic [21, 22], a measure of inter-rater reliability, from these code assignments. RH and BP discussed any responses to which their code assignments disagreed, and further refined the code definitions as needed.

The above process was repeated for each of the four PMQ probes, yielding kappa values of 0.42 for RD, 0.78 for UR, 0.65 for SMDS, and 0.70 for DMSS. We decided that the kappa value for the RD probe was unsatisfactory, and selected a different set of 40 responses for RH and BP to code separately after further refinement of the RD code definitions. That subsequent set of code assignments yielded a kappa value of 0.63, which we deemed sufficient to proceed. We note that this wide range of kappa values is indicative of the challenges in capturing the subtlety in the student thinking prompted, to varying degrees, by each probe.

RH then used our final code definitions to code responses from Spring 2017 and Spring 2018. Before coding, RH anonymized and shuffled together all of the pre and post responses to avoid unintentional bias or other systematic effects based on student identity or timing.

Here, we report a subset of results from a preliminary analysis of responses from Spring 2018, carried out by BP, as an indication of the utility of our coding scheme. Only students who completed the PMQ at both the beginning and the end of the course were included in this analysis, which resulted in a data set of 499 matched responses. We count the number of times each code was assigned in the pre, and in the post, data set, and calculate corresponding uncertainties on those counts using the binomial proportion confidence interval at the 95% confidence level. We then take the difference between the number of post responses and pre responses for each code, with propagated uncertainties. For each code, we interpret a difference with an uncertainty interval that excludes zero as a statistically significant shift between pre and post data sets.

III. RESULTS

The full set of codes in our new scheme consists of 12-16 codes for each of the four PMQ probes we studied, roughly evenly split within each probe between point paradigm codes, set paradigm codes, and unknown codes. We present a selection of codes in Table I, as examples of the new PMQ coding scheme. We selected these codes to illustrate some of the more subtle distinctions we encountered while creating our new coding scheme, which we describe in this section. We also include examples drawn from the PMQ responses in our data set to exemplify these codes.

The RD probe prompts students to decide whether to repeat a measurement several times, exactly two times, or only once. Even if a student decides that multiple trials are necessary, their explanation can still fall into the point paradigm if their reasoning evaluates each data point in isolation to decide if it is the true value of the measurand. The P2 code of the RD probe (denoted RD-P2) represents one such line of reasoning, as in the explanation, "Multiple trials help confirm results and eliminate previous errors." However, some explanations de-

TABLE I. Selected codes from the new PMQ coding scheme.

Probe	Identifier	Name	Definition: "Argument is that..."
RD	P2	Identify the outliers after all measurements	...repeated measurements are needed in order to know which measurements were mistakes or outliers, after all measurements are taken. This code includes the idea that the experimenter must get the same result at least twice for it to be correct.
	U2	More data cancels out error	...the experimenter needs to take more data to cancel or outweigh the effect of error.
UR	P1	Choose single value	...the experimenter should choose a single value to report (for any reason).
	S2	Why average is useful	...reporting the average is best, because (in general) it accounts for fluctuations or errors, or because it predicts future measurements.
	S3	Why average is appropriate in this case	...reporting the average is best because all of this data matters, or because the spread of this data is small enough. Includes reporting all data as well as the average.
	S4	Report average and spread	...the experimenter should report the average and the uncertainty/range/spread.
	S5	How to compute	...the response explains how to compute the average.
DMSS	P3	Means close enough, treats average as point	...the groups agree because the means are close enough.
	S1	Means are close enough, talks about statistical variation in general	...the groups agree because the averages are close enough. Argument contains no reference to spreads, but does discuss statistical variation in general.

scribe the purpose of multiple data points more generally, for example, "Repetition can help minimize error." The reasoning behind this second response hinges on the meaning of the word "error." The RD-U2 code was assigned in this case to capture the ambiguity between two possible interpretations: one in which the word "error" refers to mistakes, and the other in which "error" refers to statistical uncertainty, thus making the response fall into the set paradigm.

The UR probe is shown in Fig. 1. While most students at CU respond to this probe with reasoning aligned with the set paradigm [16], there are nonetheless salient differences between their responses. For example, one student reported the average because it "takes into consideration of all the points." This reasoning aligns with the set paradigm, as it sees each data point as contributing to a single result of the measurement process. We assigned the UR-S2 code to this explanation. Another student wrote, "I decided to take the average of all the results, seeing as they seem to fall within a decently confined range." While similar to the previous response, this one supports the use of an average because of features of this data set *in particular*. While still falling within the set paradigm, as it takes into account the set of measurements as a whole, it relies on these particular data having a small enough spread. The response suggests that in cases where the spread was larger, an average might not be the appropriate value to report. We assigned the UR-S3 code here, distinguishing it from the first response represented by UR-S2.

The DMSS probe asks students to decide whether two sets of data agree with each other. The probe presents two tables of five values each, and the average for each set of values. The two sets have different means, but very similar spreads, with the difference in averages being less than the standard deviation of either data set. While this probe strongly prompts students to use averages in their response, responses can still fall in the point paradigm if the underlying reasoning treats the two averages as points in their own right, considered independently as candidates for the true value of the measurand. For example, the explanation, "[The groups agree because] the difference between the two averages is very low," takes only the two average values into account, ignoring other properties of the data. We assigned the DMSS-P3 code to this response. In contrast, the explanation, "They most likely agree, although it would be easier to decide if they agree if we knew the tolerance in the measurement," recognizes that the result of a measurement involves more than just an average. We assigned the DMSS-S1 code to this response. It does not specify what the "tolerance" is, how to calculate it, or how to use it to decide if the results agree, each of which would be represented by other set paradigm codes. Nonetheless, this response recognizes that information beyond the two average values is required, and thus falls within the set paradigm as represented by the code DMSS-S1.

To illustrate the insight offered by our new coding scheme, we will focus on results using the UR probe from the pre and post administrations of the PMQ in the Spring 2018 semester. For context, in Spring 2018, of all the matched UR responses, 93.8% of pre and 97.4% of post responses aligned with the set paradigm. While these two distributions are statistically distinct (using the Mann-Whitney U-test at the 5% significance level [23]), they offer little practical significance due to the high incidence of set reasoning in pre and post.

The differences between the post and pre counts for each new UR code are displayed, along with corresponding uncertainty intervals, in Fig. 2. There were statistically significant decreases in the number of responses assigned UR-S2, UR-S3, UR-P1, and UR-S5, and there was a statistically significant increase in the number of responses assigned UR-S4. The definitions of these codes are included in Table I. While the UR-P1 code is the only one of these falling under the point paradigm, the others represent various lines of reasoning that all fall within the set paradigm. Our coding scheme distinguishes between these lines of reasoning, allowing us to characterize changes in students' arguments even when the overall paradigm of those arguments remains consistent.

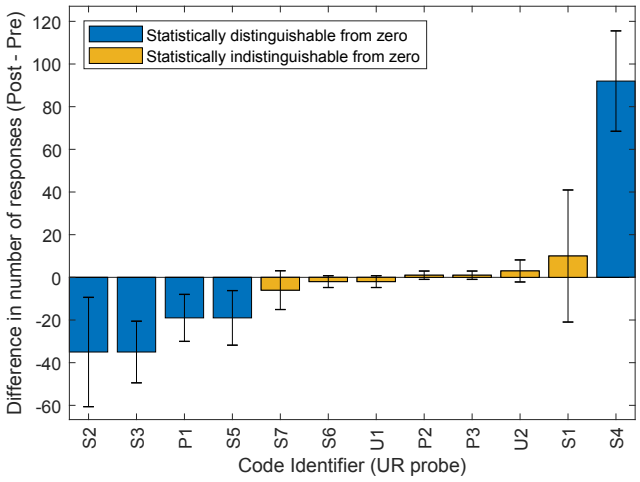

FIG. 2. Differences in pre-post code distributions from Spring 2018 for the UR probe, representing 499 pre-post matched responses.

IV. DISCUSSION

In this section, we provide an interpretation of the results shown above and discuss their wider implications. First, however, we note some limitations of our study.

This work involves student responses collected from a single institution, CU, which is a large, research-focused, highly resourced, primarily white institution of a type that is overrepresented in literature [24]. Further work, with a broader range of students and institutions, is needed to determine whether our coding scheme is applicable beyond CU. Additionally, our coding scheme is based on responses of students enrolled in an introductory physics lab class, and is intended only for measuring understanding around measurement uncertainty at the introductory undergraduate level. Using these codes to categorize responses from students at other levels of study will likely first require significant modification of the coding scheme, and possibly also the PMQ probes themselves.

As a final limitation, we note that while our new coding scheme better describes the breadth and depth of CU students' reasoning around measurement uncertainty probed by the PMQ, its overall utility is still limited by the range and depth of reasoning that the PMQ itself elicits. The probes of the PMQ concern concepts and practices around basic statistical approaches to understanding and quantifying uncertainty modeled as a stochastic process. This scope is narrow when compared to the plethora of ideas and techniques that constitute measurement uncertainty in experimental physics.

We now discuss the results presented in the previous section, starting with some broader implications from our new coding scheme. We created this new scheme for the PMQ because the coding scheme from Cape Town did not transfer to Boulder at a deeper level than the paradigm model. However, a viable set of codes emerged from CU students' responses to the PMQ, suggesting that the PMQ itself is robust enough to transfer from Cape Town to Boulder and prompt a variety of responses around measurement uncertainty in a new national and institutional context.

The UR probe in particular illustrates the utility of our coding scheme in the CU context. This probe is notable because our students provide pre responses that are predominately in the set paradigm, leaving little room for shifts towards set reasoning in the post responses [16]. Such paradigm-level results offer little insight into the student reasoning that the UR probe aims to provide. On the other hand, analyzing differences within the set paradigm using our new coding scheme provides greater insight. Specifically, we see a consolidation of response types between pre and post data sets, where a variety of set arguments (UR-S2, UR-S3, and UR-S5) become less prevalent in favor of a different type of response (UR-S4) within that same paradigm. UR-S2, UR-S3, and UR-S5 concern responses that consider only the average of a set of data. However, UR-S4 indicates reasoning that takes into account other properties of a set of data, in particular its spread. We believe that the UR-S4 code represents a more complete understanding of measurement uncertainty. That interpretation suggests that students displayed a fuller understanding of measurement uncertainty over the course of the introductory physics lab course in Spring 2018. This added value over the course of the semester, as measured by the UR probe of the PMQ, would not be apparent by analyzing PMQ responses merely using point and set paradigms, or in a more general sense, without measuring learning outcomes with the depth and specificity necessary to study learning in lab settings.

V. CONCLUSIONS

This work addresses the two goals discussed at the outset. Regarding the first goal, we described our process for developing a new coding scheme for the PMQ. The codes we created emerged from PMQ responses from students enrolled in multiple semesters of the introductory physics lab course at CU, and capture the range of responses observed in that context. Through independent coding and discussion of definitions among multiple researchers, we refined our set of codes to capture and distinguish the scope of underlying reasoning from students in that course. Regarding the second goal, we demonstrated the utility of our coding scheme by analyzing pre and post responses, and showed that the results of such analyses offer insight beyond the paradigms used previously.

Further study will involve additional analysis of these results, providing a more complete picture of how students' understanding of measurement uncertainty changed after taking the intro physics lab at CU. This insight will help to improve that course and potentially other introductory physics labs, and contribute to a better understanding of the varied and invaluable learning opportunities that lab courses offer.

ACKNOWLEDGEMENTS

Jacob T Stanley assisted in refining the initial set of PMQ codes developed at CU. Saalih Allie engaged in enlightening discussions about the PMQ. Support from NSF under Grant Nos. PHYS-1734006, DUE-1726045, DUE-1323101, DUE-1525331, and DMR-1548924. Support also from the Assoc. Dean for Educ. in CEAS and the College of A&S at CU.

[1] National Research Council, *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering* (Natl Acad Press, Washington, DC, 2012).

[2] PCAST STEM Undergraduate Education Working Group, *Engage to Excel: Producing One Million Additional College Graduates with Degrees in STEM* (Executive Office of the President, 2012).

[3] AAPT Committee on Laboratories, *AAPT Recommendations for the Undergraduate Physics Laboratory Curriculum* (Am Assoc Phys Teach, 2014).

[4] Joint Task Force on Undergraduate Physics Programs, *Phys21: Preparing Physics Students for 21st Century Careers* (Am Phys Soc, Am Assoc Phys Teach, College Park, MD, 2016).

[5] C. Walsh, K. N. Quinn, C. Wieman, and N. Holmes, Quantifying critical thinking: Development and validation of the physics lab inventory of critical thinking, Physical Review Physics Education Research **15**, 010135 (2019).

[6] B. M. Zwickl, D. Hu, N. Finkelstein, and H. J. Lewandowski, Model-based reasoning in the physics laboratory: Framework and initial results, Physical Review Physics Education Research **11**, 020113 (2015).

[7] D. R. Dounas-Frazerand H. J. Lewandowski, The Modelling Framework for Experimental Physics: description, development, and applications, European Journal of Physics **39**, 064005 (2018).

[8] B. Campbell, F. Lubben, A. Buffler, and S. Allie, Teaching scientific measurement at university : understanding students â ideas and laboratory curriculum reform, AJRMSTE , 1 (2005).

[9] J. Dayand D. Bonn, Development of the concise data processing assessment, Physical Review Physics Education Research **7**, 010114 (2011).

[10] H. Eshachand I. Kukliansky, Developing of an instrument for assessing students' data analysis skills in the undergraduate physics laboratory, Can J Phys **94**, 1205 (2016).

[11] A. Madsen, S. B. McKagan, E. C. Sayre, and C. A. Paul, Resource Letter RBAI-2: Research-based assessment instruments: Beyond physics topics, American Journal of Physics **87**, 350 (2019).

[12] F. Lubbenand R. Millar, Children's ideas about the reliability of experimental data, International Journal of Science Education **18**, 955 (1996).

[13] S. Allie, A. Buffler, B. Campbell, and F. Lubben, First-year physics students' perceptions of the quality of experimental measurements, International Journal of Science Education **20**, 447 (1998).

[14] A. Buffler, S. Allie, and F. Lubben, The development of first year physics students' ideas about measurement in terms of point and set paradigms, Int J Sci Educ **23**, 37 (2001).

[15] A. Buffler, S. Allie, F. Lubben, and B. Campbell, Evaluation of a research-based curriculum for teaching measurement in the first year physics laboratory, 4th Conference of the European Science Education Research Association **09**, 19 (2003).

[16] H. J. Lewandowski, R. Hobbs, J. T. Stanley, D. R. Dounas-Frazer, and B. Pollard, Student reasoning about measurement uncertainty in an introductory lab course, in *2017 Physics Education Research Conference Proceedings* (American Association of Physics Teachers, 2018) pp. 244–247.

[17] B. Pollard, R. Hobbs, J. T. Stanley, D. R. Dounas-Frazer, and H. J. Lewandowski, Impact of an introductory lab course on students' understanding of measurement uncertainty, in *2017 Physics Education Research Conference Proceedings* (American Association of Physics Teachers, 2018) pp. 312–315.

[18] B. Pollardand H. J. Lewandowski, Transforming a large introductory lab course: impacts on views about experimental physics, in *2018 Physics Education Research Conference Proceedings* (American Association of Physics Teachers, 2019).

[19] H. J. Lewandowski, D. R. Bolton, and B. Pollard, Initial impacts of the transformation of a large introductory lab course focused on developing experimental skills and expert epistemology, in *2018 Physics Education Research Conference Proceedings* (American Association of Physics Teachers, 2019).

[20] T. S. Volkwyn, *First year students' understanding of measurement in physics laboratory work*, Ph.D. thesis (2005).

[21] J. Cohen, A Coefficient of Agreement for Nominal Scales, Educ Psychol Meas **20**, 37 (1960).

[22] N. J. M. Blackmanand J. J. Koval, Interval estimation for Cohen's kappa as a measure of agreement, Stat Med **19**, 723 (2000).

[23] H. B. Mannand D. R. Whitney, On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other, Ann Math Stat **18**, 50 (1947).

[24] S. Kanimand X. C. Cid, The demographics of physics education research, arXiv.org **1710.02598** (2017).