HOW ARE GEOMETRIC PROOF PROBLEMS PRESENTED? CONCEPTUALIZING AND MEASURING TEACHERS' RECOGNITION OF THE DIAGRAMMATIC REGISTER

Patricio Herbst	Karl W. Kosko	Justin K. Dimmel
University of Michigan	Kent State University	University of Michigan
pgherbst@umich.edu	kkosko1@kent.edu	jkdimmel@umich.edu

We describe the development of measures of teachers' recognition of an instructional norm—that proof problems in high school geometry are presented in a diagrammatic register. A first instrument required participants to openly respond to depictions of classroom scenarios in which the norm was breached. A second instrument was a survey that required participants to rate the extent to which they agreed with various explicit statements about instruction. A third instrument capitalized on pros of the other two. We demonstrate how this instrument development process improved our conceptualization of the components included in the diagrammatic register norm.

Keywords: Geometry and Geometrical and Spatial Thinking, Instructional activities and practices, Measurement, Reasoning and Proof

We describe the process of conceptualizing a norm of instruction and measuring its recognition by teachers. The paper illustrates how the process of improving the instruments led to an improved conceptualization of the construct being measured. Weiss and Herbst (2007) had observed that while theorems are often stated in geometry textbooks using a *conceptual* register (referring to mathematical objects by their names, e.g., base angles, diagonals), proof problems often use a *diagrammatic* register, whereby the 'givens' and the 'prove' are stated in terms of specific objects in a diagram (i.e., using their labels). Moreover, diagrams may add information not stated as "givens" that is nonetheless essential to prove the conclusion (these include properties of betweenness, collinearity, concurrency, intersection, orientation, and separation). Our investigation aims at (1) refining the meaning of the statement that proof problems are presented in the diagrammatic register and (2) determining to what extent geometry teachers recognize such presentation as normative.

We build on scholarship that studies social practices and their participants' tacit knowledge (e.g., Bourdieu, 1990; Collins, 2010; Garfinkel & Sacks, 1970). This scholarship offers the notion that practices include regularities that are often tacit even though they are experienced as normative: We call those *norms* and apply that notion to the study of mathematics teaching. Earlier work on the notion of classroom norms (and germane notions such as *cultural script*) has drawn empirical support on detailed analysis of cases of teaching (e.g., Bauersfeld, 1988; Stigler & Hiebert, 1999; Yackel & Cobb, 1996) or on the analysis of practitioners' reactions to cases of instruction (e.g., Jacobs & Morita, 2003; Nachlieli & Herbst, 2009). Scholarship in social psychology (Aarts & Dijkterhuis, 2003; Nolan, et al., 2008) shows examples of how recognition of situational norms can be studied empirically and some of that methodology has percolated to the study of general norms in teaching (Hora & Anderson, 2012). This paper contributes to designing instruments to confirm quantitatively the recognition of norms of instructional situations in mathematics: By instructional situations we mean particular segments of classroom mathematical work that have a taken-as-shared exchange value among the instructional goals at stake in a given course of studies (Herbst, 2006). In this case we focus on the situation of "doing proofs" in which proofs of particular propositions exchange for students' skill at doing proofs.

We focus on the design of instruments to measure recognition of the norm that proof problems are stated in the diagrammatic register. Our initial understanding of what could be meant by *diagrammatic register* was holistic—it referred to relationships between the problem statement and the diagram included. We expected that practitioners would hold that norm tacitly: They would recognize when a proof problem had breached this norm and be able to produce proof problems that complied with the norm, but they would not necessarily be able to say what the norm consists of.

Methods

Initially, we created two kinds of instruments. One, a "tacit norm recognition instrument" (a.k.a., N1), kept the phenomenon at a holistic level, requiring participants to evaluate the work of teacher in classroom scenarios where the teacher posed a proof problem that breached the norm. Scenarios were rendered as image sequences with cartoon characters enacting teacher and students and speech bubbles for their talk. The open ended questions asked participants to describe what they saw happening and to evaluate the teacher's work facilitating the doing of a proof. Four different scenarios were presented to participants (item sets 21002, 21003, 21004, 21006), each of which enacted a breach of the norm in different ways. For example in one scenario the teacher described in writing the figure to which the proof problem referred but did not draw a diagram for it. Participants were told that the scenarios were about doing proofs in geometry but were not told that they included breaches of norms or that there was anything special about the handling of diagrams.

The second instrument, an "explicit norms recognition instrument" (a.k.a., N2), required participants to rate statements that described, in general, possible behaviors of a teacher posing proof problems. Participants had to rate how appropriate and how typical were actions described in statements, such as: "the teacher provides a diagram for students to use while doing a proof". The creation of this instrument gave us a first opportunity of laying out in a more analytic fashion what the "diagrammatic register" norm consisted of. In doing so we proposed five distinct subnorms making up the diagrammatic register. Concisely stated, the initial statements of these subnorms were as follows (construct IDs in brackets): (1) the teacher provides a diagram that has the givens marked when assigning a proof [DP21]; (2) the teacher provides a diagram for students to use while doing the proof [DP24]; (3) the points of the diagram which are needed for the proof (though not necessarily all of the points) are provided and labeled by the teacher [DP36]; (4) the statement of the proof problem uses symbols and labels for the elements of the diagram (e.g., $AB \perp CD$) [DP39]; (5) the diagram the teacher provides accurately represents the concepts at stake in the proof [DP41].

To examine responses to N1 we created a coding scheme that assessed open responses for evidence that participants recognized breaches of the diagrammatic register norm and for evidence of whether participants appraised the teaching negatively. Reliability was established by having two independent coders code through responses and resolve disagreements on a case-by-case basis. To examine internal reliability of constructs in the N2 instrument, we implemented classical item analysis (CIA; see Crocker & Algina, 2006).

Data Sources and Analysis

The N1 instrument was piloted with 50 teachers including 33 that had three or more years of experience teaching geometry (EGT) and 17 teachers without such experience (thus including novices as well as teachers experienced teaching other courses). The N2 instrument was piloted with 44 teachers including 28 EGT and 16 teachers without such experience. Table 1 shows how many participants recognized the breaching of the diagrammatic register norm in each of the N1 stories. Across items, experienced geometry teachers had higher totals for recognition of the norm. Binomial tests, assuming a null hypothesis that recognition and nonrecognition are equally likely, showed statistically significant results for experienced geometry teachers in story 21002 (p<.05), and 21003 (p<.05), while for these same items the other participants showed no significant difference from chance in their rate of recognition of the norm.

Table 1*: Number of teachers who recognized breach of the norm.*

	21002	21003	21004	21006
Exp. geometry teachers (>3 years)	18	18	11	5
Other teachers	7	10	4	3
Total	25	28	15	8

Responses were also coded for the presence of negative appraisals (Martin & White, 2007) of the depicted teaching. We defined the variable "discomfort with the scenario" to capture participants' critiques not sufficiently focused to count as recognitions of the norm. For example, "discomfort" captured entries where participants critiqued that the teacher "did not write anything down." "Discomfort" and "norm recognition" are mutually exclusive—only participants that did not recognize the norm were coded for having discomfort with the scenario. Table 2 shows the aggregate of those participants that either recognized the norm or registered discomfort with the teaching.

Table 2*: Aggregate proportion of participants that recognized the norm and participants that registered some discomfort with each scenario.*

	21002	21003	21004	21006
Exp. geometry teachers (>3 yrs.)	30	29	17	24
Other teachers	10	14	6	14
Total	40	43	23	38

The data suggests that participants may be reacting negatively to the scenario in ways that are not localized around the breach of the norm and the specific features of the diagrammatic register. Indeed, even among participants that did recognize the norm, that recognition was not always focused on features commensurate with the subnorms identified above. Those results were not surprising on account of our hypothesis that the diagrammatic register norm is tacit: Unlike other norms in "doing proofs"—such as the norm that every statement must be justified by a reason (Nachlieli & Herbst, 2009)—the norm that problems are presented in the diagrammatic register is not explicitly addressed in the work of teaching. Earlier work in ethnomethodology suggested that when such implicit norms are breached, participants engage in repair strategies—ways of signaling that something is amiss—though these strategies do not always include pointing at what was amiss (e.g., reframing the activity as a case of a different situation).

This lack of specific reference to the features of the register within the open responses was the principal motivation that led us to conceive of a new instrument (N3, described below) to study the diagrammatic register. It seemed that, since the subnorms of the diagrammatic register are tacitly held, testing for these subnorms via open response prompts placed too onerous a reporting burden on the participant. If participants did not comment on specifics of how problems were stated this could mean that participants did not know how to focus their reaction—an explanation that coheres with the hypothesis that the norm is tacit. The other reason that called for developing N3 came from the analysis of pilot data on N2.

Participants made two ratings of each of the N2 statements: An appropriateness rating—this was a 6-valued rating scale, ranging from "Very Inappropriate" to "Very Appropriate"—and a typicality rating—this was a 4-valued rating scale that ranged from "It Never or Hardly Ever Happens" to "It Always or Almost Always Happens". The two kinds of questions aimed at complementary aspects of normativity (appropriateness and frequency) so we examined their internal reliability separately.

Table 3*: Reliability for the N2 Constructs.*

Construct	Reliability for "Appropriateness" Scale	Reliability for "Typicality" Scale
DP21 [Givens marked]	$\alpha = .78$	$\alpha = .61$
DP24 [Diagram provided]	$\alpha = -.02$	$\alpha = -.06$
DP36 [Labeled points]	$\alpha = .50$	$\alpha = .70$
DP39 [Statement uses labels]	$\alpha = -.37$	$\alpha = .15$
DP41 [Diagram Accuracy]	$r = .40$	$r = .33$

Table 3 shows results of the item analysis for the responses to the N2 item, indicating reliability for some constructs (DP21, DP36, DP41) but not for others (DP24, DP39). DP41 had only two items, so while alpha was not calculated, responses to both items showed moderate correlations. For DP39, items generally performed poorly unless recoded to assume the norm was that all points had to be labeled, and not just the ones relevant for a proof (in which the alphas improved to over .70 for each scale).

Table 4*: Correlations with Experience Teaching Geometry for Appropriateness and Typicality.*

	DP21		DP24		DP36		DP39		DP41	
	App	Typ	App	Typ	App	Typ	App	Typ	App	Typ
Taught >3 years Geometry	-.04	.05	.03	.08	-.36*	-.26[a]	.08	.20	-.09	-.05
Taught ≥ 1 year Geometry	-.27[a]	.12	.02	-.12	-.36*	.19	.07	.23	-.19	.12

[a] *$p < .10$, *$p < .05$, **$p < .01$*

Note: 'App' and 'Typ' represent correlations for the appropriateness and typicality scales, respectively.

Striving to better understand why the constructs performed the way they did, we looked at how participants answered items, by correlating their status of being an experienced geometry teacher (or not), as well as if they had ever taught geometry (or not) with the construct scores. These results, shown in Table 4, showed few correlations between the constructs and geometry teaching experience. The only statistically significant correlations were negative, suggesting that participants rejected the explicitly stated norms.

Discussion: Revisions Needed to the Instruments

Several issues were raised by the analysis of the data from the first pilot. First, the results of the first pilot indicated that the N1 instrument was insufficiently sharp: It did not elicit commentary from participants that pointed specifically to the features of the diagrammatic register norm. Another issue with the N1 instrument was that some of the discomfort reported by participants could be attributed to ancillary features of the scenarios—e.g., that the teacher had not written the statement of the problem on the board. In revising scenarios for a second pilot we avoided those confounding elements, but our review of the results from the first pilot also raised issues with the N2 instrument. Specifically, the item analysis from N2 led us to question whether recognition of the diagrammatic register subnorms could be assessed with instruments that asked participants to rate general statements. We also realized that in looking for explicit general statements of the subnorms of the diagrammatic register we had failed to include any indications of what mathematical properties are often communicated through the diagram. In the case of developing scenarios to breach the diagrammatic register norm for the N1 instrument, we had assumed that we should choose problems that involved collinearity, separation, concurrency, or betweenness, and that we should present those problems without using diagrams. But we had failed to include any N2 item that explicitly tested for recognition that such properties are normatively telegraphed by the diagram. As a result, we had created some tacit items (for the N1 instrument) that seemed to lose part of their meaning when made into explicit general statements (for the N2 instrument). The logistical and technical difficulties we had run into when examining the pilot data had led us to realize that there were gaps in our conceptualization of what the subnorms really should be. We had the chance to fill those gaps when we conceptualized a new instrument, one that we refer to as N3.

Table 5*: Revised Subnorms of Diagrammatics Register After Initial Piloting.*

Designation	Subnorm statement
Subnorm 1 (New)	The statement of the problem does not make explicit properties of betweenness, intersection, separation, collinearity, or concurrency, which are left for the diagram to communicate.
Subnorm 2 DP21	The teacher provides a diagram for students to use while doing the proof.
Subnorm 3 DP39	The teacher assigns a proof problem with an accompanying diagram where the points needed in the proof are labeled (but not necessarily all points).
Subnorm 4 DP24	The proof problem is stated using symbols and labels for elements of a diagram (e.g., $AB \perp CD$).
Subnorm 5 DP41	When a teacher provides a diagram accompanying a proof problem, the diagram is accurate.

We drew an analogy to an optometrist trying different lenses on a patient to design a novel format for an item that could blend the contextualization virtues of the N1 instrument (where participants handle a specific proof problem) and the analytic virtues of the N2 instrument (where it is possible to expose participants to many items and thus be able to do an item analysis). The resulting N3 instrument asks participants "which of the two proof problems below is more appropriate for geometry teachers to present to their students" and offers two problems that differ from each other only in regard to their compliance with or breach of one subnorm of the diagrammatic register. The participant answers using a 6-point scale, with options that include "Option A is much more appropriate" (viz., somewhat more appropriate, slightly more appropriate), and same options for Option B. To concentrate on the norms that had been more problematic when developing N2, we created N3 items that tested the subnorms listed in Table 5.

Table 6. *Reliability Measures for Revised Instruments for N2 & N3.*

Assessed Norms	Reliability (α) for N2 in 2nd Pilot	Reliability (α) for N3 in 2nd Pilot
Subnorm 1: The statement of the problem does not make explicit properties of order, separation, collinearity, or concurrency, which are left to the diagram to communicate.	-.10	.64
Subnorm 2: The teacher provides a diagram for students to use while doing the proof.	.69	.80
Subnorm 3: The teacher assigns a proof problem with an accompanying diagram where the points needed in the proof are labeled (but not necessarily all points).	-.28[1] .61[2]	-.12[1] .69[2]
Subnorm 4: The proof problem is stated using symbols and labels for elements of a diagram (e.g., $AB \perp CD$).	.63	.77
Subnorm 5: When a teacher provides a diagram accompanying a proof problem, this diagram is accurate.	.55	.82

[1]*Assumes the norm of "only points relevant to completing the proof are labeled."*
[2]*Assumes the norm of "all points on the diagram are labeled."*

This N3 instrument, along with a revised version of the N2 instrument (whose revisions were a consequence of the further articulation of the subnorms described above), was piloted during three data collection sessions in May and June of 2012. Forty-nine participants completed the revised N2 instrument, while 52 participants completed the N3 instrument. This second round of piloting showed higher validation scores than in the first pilot, as reported in Table 6. Still, while the data showed improvement in the reliabilities for four of the subnorms in the revised N2 compared to its first version, N2 was still not effective assessing recognition of the first subnorm. Overall, while not completely successful, N3 is providing more reliable estimations of participants' recognition of the subnorms that constitute the diagrammatic register. (Incidentally, note that in the second pilot we did not test for recognition of DP 21, the norm that givens are marked, because that one had achieved acceptable alpha levels in the first pilot.)

The N1 instrument was also revised to sharpen the scenarios. One scenario (21004) was withdrawn and two new ones were included (21005 and 21007). Additionally, the questions in each item set were revised. For each scenario, participants were first asked to describe what they saw happening. Then they were asked to rate, on a 6-point scale, how appropriate the teacher's facilitation of the work on a proof was and they were given a box to explain their rating. Finally they were asked to rate, on a 6-point scale, how appropriate the description of the proof problem assigned to the class was and they were given a box to explain their rating. As in the case of the first pilot, we coded the open response questions for evidence of recognition of a breach of the norm and for evidence of discomfort with the scenario. We defined the aggregates "Norm recognition" (NR) and "Repair" (RP) as follows: For each item set *j*, participants were assigned a 1 for NR(*j*) if at least one of the open responses contained evidence that the participant had noticed a breach in a specific subnorm of the diagrammatic register; participants were assigned a 1 for RP(*j*) if they had a 1 for NR or if in at least one of the open responses they indicated discomfort with the teaching represented in the scenario. Reliability for this coding was very good, as attested by kappa values reported in Table 7. The table also contains basic counts of these variables for the whole sample, along with their significance, assessed using binomial tests against a null hypothesis that recognition or non-recognition of a breach in or a repair to the scenario occurred with equal probability.

Table 7: NR(*j*) and RP(*j*) totals per session for second N1 pilot study, with kappa (κ) statistics for reliability, and with binomial probabilities.

Session		NR(*j*)			RP(*j*)		
	n	Total	κ	*p*	Total	κ	*p*
21002	42	32	0.77**	0.0003	37	0.82**	<0.001
21003	39	23	0.78*	0.0686	29	0.77**	<0.001
21005	39	22	0.87*	0.0928	34	0.94**	<0.001
21006	40	17	0.86*	0.0689	30	0.95**	0.001
21007	42	18	0.95*	0.0804	28	0.86**	0.012

**Indicates results that are significant at the .10 level*
***Indicate results that are significant at the .05 level*

These results of the binomial tests reported in Table 7 show that overall participants recognized the breach of the tacit norm in these items at a probability that differed significantly from what would be expected from chance alone. In this respect, item 21002 is set apart from the other breaching items. One reason that could explain this difference is that the classroom story prompt for item 21002 depicted the teacher breaching the tacit norm without an apparent reason for having done so. Whereas in the other items, while the depicted teacher still breaches the norm, in each of these instances there were—by design—mitigating factors suggested by the scenario that could have accounted for the teacher's departure from the norm, such as a teacher's wish to emphasize the conceptual (rather than diagrammatic) statements of particular geometry concepts. The fact that participants offered repairs to the scenario for each of the breach items suggests that even though participants were not as likely to recognize a specific breach in items 21003-21007, they still noticed that something was amiss in some way. Thus, it seems that this set of items is effective detecting teachers' recognition of the diagrammatic register norm.

Conclusion

The development of instruments to measure teachers' recognition of an instructional norm moves our work forward in the study of a phenomenon that is specific to mathematics teaching—i.e., specific to the work of teaching, and specific to the mathematics at play. We contend that this work is also valuable insofar as it illustrates a process of discovery in our field: A dialectics of conceptualization and measurement in the research process that challenges the received wisdom whereby small scale, qualitative, exploratory studies provide sufficient material for theoretical conceptualization while larger scale, quantitative studies are needed only eventually and just to verify or falsify theoretical assertions. Consistent with contemporary views on measurement (e.g., Wilson, 2005), the study shows how the goal of measurement and the analysis of a relatively large set of responses to initial instruments led to further conceptualization of the theoretical construct we wanted to confirm, revise the instruments to actually measure this construct, and obtain results that allow us to report on the extent to which the hypothesized construct was present.

Acknowledgments

Work reported here was done with the support of National Science Foundation grant DRL-0918425 to P. Herbst and D. Chazan. All opinions are those of the authors and do not necessarily represent the views of the Foundation.

References

Aarts, H. & Dijkterhuis, A. (2003). The Silence of the Library: Environment, Situational Norm, and Social Behavior. *Journal of Personality and Social Psychology, 84*(1), 18–28

Bauersfeld, H. (1988). Interaction, construction and knowledge. In D. Grouws and T. Cooney (eds.), *Effective Mathematics Teaching*, Lawrence Erlbaum, Hillsdale, New Jersey.

Bourdieu, P. (1990). *The logic of practice*. Stanford, CA: Stanford University Press.

Collins, H. M. (2010). *Tacit and explicit knowledge*. Chicago: The University of Chicago Press.

Crocker, L, & Algina, J. (2006). *Introduction to classical and modern test theory*. Mason, OH: Wadsworth.

Garfinkel, H. and Sacks, H. (1970). On Formal Structures of Practical Action. In J. McKinney and E. Tiryakian (eds.) *Theoretical Sociology*. New York: Appleton-Century-Crofts, pp. 337-366.

Herbst, P. (2006). Teaching geometry with problems: Negotiating instructional situations and mathematical tasks. *Journal for Research in Mathematics Education*, 37, 313-347.

Hora, M. & Anderson, C. (2012). Perceived norms for interactive teaching and their relationship to instructional decision-making a mixed methods study. *Higher Education: The International Journal of Higher Education Research*. DOI 10.1007/s10734-012-9513-8.

Jacobs, J. and Morita, E. (2002). Japanese and American teachers' evaluations of videotaped mathematics lessons. *Journal for Research in Mathematics Education,* 33, 154-175.

Martin, J., & White, P. (2007). *The language of evaluation: Appraisal in English*. NY: Palgrave Macmillan.

Nolan, J., Schultz, P. W., Cialdini, R., Goldstein, N., and Griskevicius, V. (2008). Normative Social Influence is Underdetected. *Personality and Social Psychology Bulletin, 34*(7), 913-923.

Stigler, J.W., & Hiebert, J. (1999). *The teaching gap: Best ideas from the world's teachers for improving education in the classroom*. New York: Free Press.

Weiss, M. and Herbst, P. (2007, April). "*Every single little proof they do, you could call it a theorem*"*: Translation between abstract concepts and concrete objects in the Geometry classroom*. Paper presented at the annual meeting of AERA. Chicago.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. New York: Psychology Press.

Yackel, E., & Cobb, P. (1996). Sociomathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education*, 27, 458-477.