The affordances and challenges of using comparative longitudinal designs in matter LP research: Lessons from the Inquiry Project[1]

Carol L. Smith
Department of Psychology, University of Massachusetts/Boston

Nathaniel J. S. Brown
Educational Research, Measurement, and Evaluation, Boston College

Paper presented at NARST Symposium: Methodological Trade-offs in LP Work
April 2, 2014

As described by Rivet and Duncan in their opening commentary, learning progression researchers widely share the assumption that a learning progression is a research-based proposal for how ideas about a content domain could coherently evolve over a long period of time, given appropriate instruction, to bridge between a lower and an upper anchor (Corcoran, Mosher, Rogat, 2009). Further, they seek to identify those "big" disciplinary ideas and inquiry practices that may be most foundational and generative for further learning, and argue that one needs to assess knowledge not as isolated declarative propositions that can be memorized and repeated, but as activated and used in meaningful problems to predict, explain, and understand events. This has led to the powerful idea of considering the different kinds of learning performance that can be used to assess understanding of a big idea – for example, using the idea in making a prediction, explanation, argument, etc. (Krajcik, MacNeil, & Reiser, 2008).

At the same time, there are interesting differences in the methods used to develop and validate LPs by different researchers, which may stem (in part) from differences in how they view *what* is progressing in an LP as well as the age group of those they are studying. In our paper, we'd like to highlight how members of the Inquiry project conceptualize LPs and consider how that influenced their choice of methods. We will also consider the affordances and challenges of these methods, along with the special challenges when one is working with younger elementary school children, as some of these issues have not been as widely debated and discussed.

In their recent matter LP work, done in conjunction with their colleagues Sue Doubler and David Carraher at TERC who spearheaded the Inquiry Project, Smith and Wiser argued that "what" progresses is a complex knowledge network comprising multiple inter-related knowledge elements (concepts, beliefs, models, practices) and that bridging between the lower and upper anchor requires *a series of broad reconceptualizations* because the concepts, beliefs, models, and practices of kindergartners are fundamentally different from those of scientists (Wiser & Smith, 2013). These reconceptualizations center on the construction of new models that coherently organize and represent relations among a set of concepts, embody new ontological, mathematical and epistemological commitments, and represent an explanatory "advance" over earlier models. Each reconceptualization also brings students "conceptually closer" to the scientists' views, thus serving as a productive stepping stone for further learning. A central task for LP researchers is to identify and characterize the diverse elements and organizing principles of each of these stepping

[1] The Inquiry Project was supported by DRK-12 NSF grant #0628245, for which Susan Doubler, David Carraher, Jodi Asbell-Clarke (TERC) and Roger Tobin (Tufts) were PIs. Senior Researchers on the Project were David Carraher, Analucia Schliemann, Carol Smith and Marianne Wiser. Curriculum Designers were Sue Doubler, Sally Crissman, Nick Hadad, and Sara Lacey. For more information about the Inquiry Project, including its curricular materials go to: http://inquiryproject.terc.edu.

stones that serve as important "intermediate" targets of instruction, and test whether organizing curricula around such targets enhances achievement of next steps in the progression.

In their work, the Inquiry Project has elaborated an *early* portion of a matter LP for grades 3-5 students (Wiser, Smith, & Doubler, 2012), guided by research that has found some difficulties older students have with the atomic molecular theory stem from their having incompatible *macroscopi*c understanding of materials and weight (e.g., if they assume all stuff is tangible, and tiny things weigh nothing, how can atoms by the constituents of matter?). Thus, a first step should be helping students construct a *compositional mental model* in which objects are seen as composed of materials that retain their identity and some properties with decomposition, even tiny pieces have weight and take up space, and materials vary in their "heaviness for size." This (continuous) macroscopic model of matter can first be developed in the context of solid, then liquid and granular materials, before being reconceptualized as a particulate model that is then generalized to materials in solid, liquid, and gaseous form.

This way of thinking about LPs stems from Smith and Wiser's life long interest in studying and understanding the process of conceptual change. Lexicalized concepts (such as material, weight, matter, density, and volume) aren't hard little atoms that maintain their identity and structure regardless of relations with other concepts. Rather they are part of complex knowledge systems and as such both shape and are shaped by those relations. That is, concepts both *participate* in beliefs and are *constituted* by them (see Amin, Smith, & Wiser, in press for more extensive discussion of this issue). For example, the belief "This tiny piece of stuff weighs nothing at all" is formulated and evaluated over one's existing concepts of stuff, weight, and number and amount; and these concepts are shaped by *othe*r beliefs in which they participate (e.g., "Weight is determined hefting", "Stuff can be seen, felt, touched", "Our senses provide true information about the world", "Amounts (and numbers) have limited granularity – e.g., there are no (or only one or two) numbers between 0 and 1." Thus, how one thinks about weight depends upon its relation to other elements in the network - what types of things have weight (e.g., objects, materials), whether and how can measure weight (by hefting, by balance scales), the range of values weight can take (e.g., integer and fractional values), the physical phenomena one connects to weight. This means that as new phenomena and relations are considered, both the concepts and the knowledge network of which they are a part can continuously grow and change. Anomalies will be encountered (e.g., how can objects be the same size but have such different weights) that call for major adjustments to the network, including re-analyzing the core properties of aconcepts, differentiating between concepts (e.g., weight of objects and density of materials), and coalescencing others (e.g., recognizing fundamental similarities between solids and liquids as forms of matter).

This view of LPs as involving a series of reconceptualizations, each of which prepares students for the next reconceptualization, has important implications not only for the design of curricula and assessments, but also the type of studies that are needed to develop and validate the LP itself. Elsewhere Smith and her colleagues have discussed in more detail how their LP framework influenced the design of the Inquiry Curriculum (Wiser, Smith, & Doubler, 2012). Briefly, they emphasized three points:

- It's important to identify and develop important **precursor ideas**, not just final form ideas. Many curricula overlook the importance of weight, material, and heaviness of material as important precursor ideas for developing elementary school students' explicit

concept of matter. The Inquiry curricula, in contrast targeted the development of these ideas.

- It is important for curricula to focus on **relations among concepts** that are revisited in different contexts and across grades. Many curricula focus on topics in isolation, rather than highlighting their inter-relations. They also don't revisit in different contexts across successive grades. Rather students may consider a topic in one grade and then not return to related topics until many grades later. In contrast, the Inquiry curricula progressively revisited and broadened the contexts considered across three grades (grades 3, 4, and 5).
- It's important to think about **productive sequences** for working on relations and revisiting concepts that preserve sense making. Many curricula may introduce a topic such as the water cycle too early before foundational ideas are in place to make sense and explain, and hence disrupt sense making. The Inquiry curricula tried to introduce new ideas so that they would be "within" reach for students at a given time, hence more likely to preserve sense making, while also providing an opportunity to "stretch" their understanding.

In our paper, however, we'd like to focus on how this view of LPs influenced the design of the assessments used by the Inquiry Project and the design of their initial pioneering study to test and examine the validity of the LP. Unlike other projects, they did not first undertake the time consuming work of validating assessments *prior* to the development of the curricular units, nor did they work with written assessments. Rather, they designed an extensive multi-part *individual structured interview* (greatly informed by prior research findings in the area) that probed different facets of understanding that they thought would be needed to develop a robust compositional model of materials and that would contribute to developing an understanding of solids, liquids, and gases as different phases of matter. These included not only tasks that probed students' understanding of relevant physical concepts themselves (e.g., material, amount of material, weight, heaviness of material, volume, matter, atoms) but also relevant mathematical ideas (rational number, fraction, repeated division, proportion) that contribute to their understanding of measurement and reconceptualization of physical quantities as dense linear continua. Given the number of ideas involved, the interview was much more extensive than normal assessments, taking on average 2 hours to complete and was typically broken into two one-hour sessions. Table 1 gives an overview of the main ideas probed in the interview, along with some sample questions. This interview was designed prior to the development of the curricular units and was given at multiple times throughout the intervention (as repeated measures). In that sense, it functioned to assess the development of broad conceptual structures, rather than a specific assessment of what was learned in particular curricular units.

Table 1: List of Key Ideas Probed in the Interview

Key Ideas	Representative questions
1-Material identity is preserved across grinding and melting	If you grind up wood, is it still wood? Will it still burn? If you file iron, is the result iron? If you melt butter, is it still butter?
2-Amount of material, weight, and balance are invariant across shape changes	[Two "identical" balls of clay. One is deformed into a pancake.] Do the ball and pancake have the same amount of clay? Weigh the same? Do they balance? How do you know?
3-Tiny (visible/invisible) pieces take up space and have weight	Does a tiny speck of clay have weight? Take up space? Could they be a piece of clay too small to see? Would it take up space? Have weight?
4-Volume is differentiated from	*Measuring (using tiles, cubes*): How much space do these cards

area in measurement and from weight in predicting displacement	cover on the table? How much space do these blocks fill up? *Water displacement* (for equal size brass and aluminum cylinders): How high will each make the water level rise?
5-Heaviness of kind of material is differentiated from and coordinated with weight and size in explanations, judgments, and inference making about objects and materials	*Explaining weight*: How can a smaller object be the heavier one? How can different size objects have the same weight? *Judging heaviness of material*: Which is made of heavier material: a copper shaving or a block of aluminum? *Inferring material*: Can you tell which of the small covered cylinders is made of the same material as this large one?
6-Solids, liquids, gases are all forms of matter and differentiated from non-matter	Which of the following are matter? Are not matter? Are you unsure about? [Child sorts 14 items, including wood, dream, water, dog, sand, air, heat, shadow] How do you know?
7-Atoms and molecules are constituents of all matter	Have you heard of atoms? Molecules? What do you think they are? How similar & different? Does this rock have atoms? If all the atoms were removed, would there be anything left?
8-There are a lot or an infinite number of numbers between any two integers.	Are there any numbers between 4 and 5? How many? If you divide 1 by 2, what do you get? Can you divide that number by 2? Can you keep going? Would you ever get to 0? Why or why not?
9-The sweetness of sugar-water mixtures depends upon proportional relations between amount of sugar and water.	Is a mixture of 2 sugar cubes in 4 cups of water sweeter than 2 cubes in 6 cups? How about 1 sugar cube in 3 cups vs. 2 sugar cubes in 6 cups? 3 sugar cubes in 8 cups vs. 2 sugar cubes in 4 cups of water? How do you know?

Clearly, the multi-part structure of the interview directly relates to their assumption that reorganizations involve changing the relations among many concepts, as well as the introduction of new concepts (through conceptual differentiations and coalescences). They chose to use structured interviews in their study (despite its time consuming nature and cost) for four main reasons. First, evidence of validity based on response processes is likely to be high, because the interviewer can check that the student is engaged and understands the question as intended (and rephrase questions as needed). Second, because the interviewer can probe for student reasoning, unexpected outcomes or responses can be detected. Third, given that they were assessing an early reorganization in a matter LP (in which concepts move from being more centered in immediate perceptual experiences to being more centered in a network of relations and measurement), it was important to create tasks that called for the manipulation and use of many physical props and to probe children's judgments and reasoning about those concrete situations. This not only makes the tasks more interesting and engaging for younger children, but also allowed them to probe relevant precursor ideas children might have before they have knowledge of specialized vocabulary. Finally, because they were working with elementary school children who were still gaining competency and mastery of reading and writing, it was important that children only needed to talk about their ideas without tediously having to write down their thinking. Obviously for younger children, reading and writing is itself a difficult (and time consuming) task that can limit how much they write, and different levels of reading and writing ability among the students would contribute to construct-irrelevant variance in the results.

Of course, there are potential threats to validity in clinical interviews and repeated measures designs as well, such as subtle interviewer bias or suggestion, and the potential for different phrasings or follow-up questions to create slightly different items for different students (for interviews), and learning from the interview itself (for repeated measures designs). Nonetheless, they tried to minimize these sources of bias through both very thorough and careful scripting of

the interview itself and training of the interviewers to be neutral. It should be noted that it was not possible to keep interviewers completely blind as to treatment status or grade level of the students (after all they went to pick them up in individual classrooms). However, most of the interviewers were not knowledgeable about the specific hypotheses of the study, and said that they generally didn't remember or think about who they were interviewing. In addition, they typically did not interview the same child over successive grades. Thus, any one child was likely to have a variety of interviewers over the course of the study. Finally to guard against learning effects from the interview itself, the tasks, although interactive, were not designed to give students feedback on "the correctness" of their answers. For example, when they had to make predictions about what would happen in different situations, such as water displacement, they were not allowed to test them out. Rather, they just were asked to justify them.

Finally, their design called for a comparative *longitudinal* teaching study. In their view, innovative longitudinal studies are needed as part of the process of developing and testing LPs because, by hypothesis, productive stepping stones are not widely fostered by existing instruction, have multiple inter-related components, and take time to construct. One cannot study the process of construction of these understandings or their impact on future learning without arranging for the conditions that support their development. In this regard, their approach is more similar to the one used by Lehrer and colleagues in developing their statistics learning progression (Lehrer, Kim, Ayers, & Wilson, in press) than the approach taken by Anderson and his colleagues in developing their learning progression for the carbon cycle (discussed by Draney et al in paper 2, where they worked on developing their assessments first) before designing their interventions. It should be noted that the Lehrer et al work on developing an LP for statistical reasoning involved an iterative series of teaching studies, over a span of 20 years.

The approach of the Inquiry Project is different from many others who have done teaching studies by its *multi-year* focus (they studied learning across three years, in grades 3 to 5) and by its *comparative design*. Comparative teaching studies are needed to test whether innovative approaches are more effective than standard instruction in producing these changes and better prepare students for next steps in learning. Such studies also allow one to investigate the similarities and differences in patterns of relations among understandings in different instructional groups. A finding of *similar patterns of relations* among component understandings in both instructional groups, even though overall levels of achievement may widely differ, would argue for important conceptual inter-dependencies that any curricular approach needs to support to be successful. It may also highlight the need for work on certain ideas that might be overlooked within more traditional curricular approaches. Different patterns of relations might signal ways that instruction itself has altered the conceptual landscape for students (e.g., by creating more robust and integrated understandings).

In summary, two groups of students were followed longitudinally from Grades 3 through 5 using repeated-measures, quasi-experimental design:

- **Treatment students** (those who received the Inquiry science curriculum for nine weeks in each of Grades 3, 4, and 5) were interviewed on four occasions over two and one-half years: (a) early Grade 3, before the Inquiry Curriculum; (b) end of Grade 3 after the first Inquiry Curriculum unit; (c) end of Grade 4 after the second Inquiry Curriculum unit; and

(d) end of Grade five after the third Inquiry Curriculum unit. These students are from five classrooms in two different schools.

- **Control students** (students from the same school who received the standard science classroom instruction and the same teachers in Grades 3-5) are interviewed at three occasions: (a) end of Grade 3, (b) end of Grade 4, and (c) end of Grade 5.

Overall, they conducted 346 interviews with between 54-67 treatment and 35-38 Control students at each occasion of testing, interviewing the same students at each occasion as much as possible. Because some students left the school across the grades or because some students did not return permission slips at a particular testing time, there was some attrition of the original sample across Grades. To maintain the same size they added new students to the Control or Treatment group for interviews among those who returned permission slips. However, students were added to the Treatment group for interviews only if they had been in the school for the duration of the study, so had experienced the full Inquiry curriculum.

Each interview followed a detailed written script, and the interviewer circled the judgment students made, wrote down student explanations, and took notes to describe specific approaches to problem solving during the interview. Interviews were also videotaped. Each interviewer then reviewed the videotape and their written notes as they entered data from the interviews into a Filemaker database for later analysis.

Smith and her colleagues then analyzed and scored the interview data in multiple ways: e.g., in terms of judgment patterns across a variety of questions, in terms of justification, and in the case of measurement tasks the actual invariant (length, perimeter, area) the student measured. In creating coding categories and coding data, they were blind to the treatment status and grade level of students. In cases where justifications or measurement approaches were analyzed, two coders scored the data independently to make sure the categories could be reliably scored (i.e., greater than 85% reliability). Otherwise, coding categories were collapsed. More details about their methods and findings are described in the final report of the Inquiry Project (Doubler et al, 2011).

The main goal of these analyses was to understand the relative difficulty of each the target understandings for grade 3 to 5 children, how student understanding of these ideas unfolded over time, and whether the Inquiry Curriculum was more successful in promoting the diverse network of ideas needed for a sound macroscopic understanding of matter than the standard science curriculum already in place in these schools. For each task, patterns of response and/or justification to individual questions were identified that indicated that students had achieved a certain benchmark understanding (in that task), and then compared the achievements of those in the Inquiry Curricula with the Inquiry Project at any given grade (using chi square tests of comparison). These analyses of the data supported the following claims:

- Elementary school children start with radically different (more perception centered) knowledge networks for thinking about matter, but can make significant progress in restructuring these ideas with appropriate curricular support. More specifically, students who had the Inquiry Curriculum made significant progress in developing *all understandings* underlying a compositional model of materials (about material, amount

of material, weight, volume, heaviness of kind of material, and matter) over the span of three years.

- This reconceptualization of their matter knowledge network takes time, with different parts having different growth trajectories (see Figure 1, for three sample items). For example, large changes in understanding that tiny (visible) pieces have weight and take up space occurred immediately by the grade 3 posttest (from less than 10% of to over 60%), then dipped a little in grade 4, before increasing again to almost 80% by the end of grade 5. There were also immediate improvements in student understanding that material identity remains invariant across decomposition and that weight of an object is not affected by shape change (although understanding of these ideas started much higher). In contrast, students made slower, but steady progress in differentiating weight and density across grades 3, 4, and 5 (as well as differentiating volume from area). Finally, progress in developing an explicit concept of matter that included solids, liquids, and gases was slower still, with the greatest change coming in grade 5. This suggests the process of reconceptualizing existing concepts (such as weight and material) is somewhat easier than introducing new concepts (density, volume, matter) via differentiation and coalescence, and may even "prepare the ground" for these later changes. It also highlights why revisiting key ideas in further contexts across adjacent grades is so important, as progress on those concepts cumulates across grades.

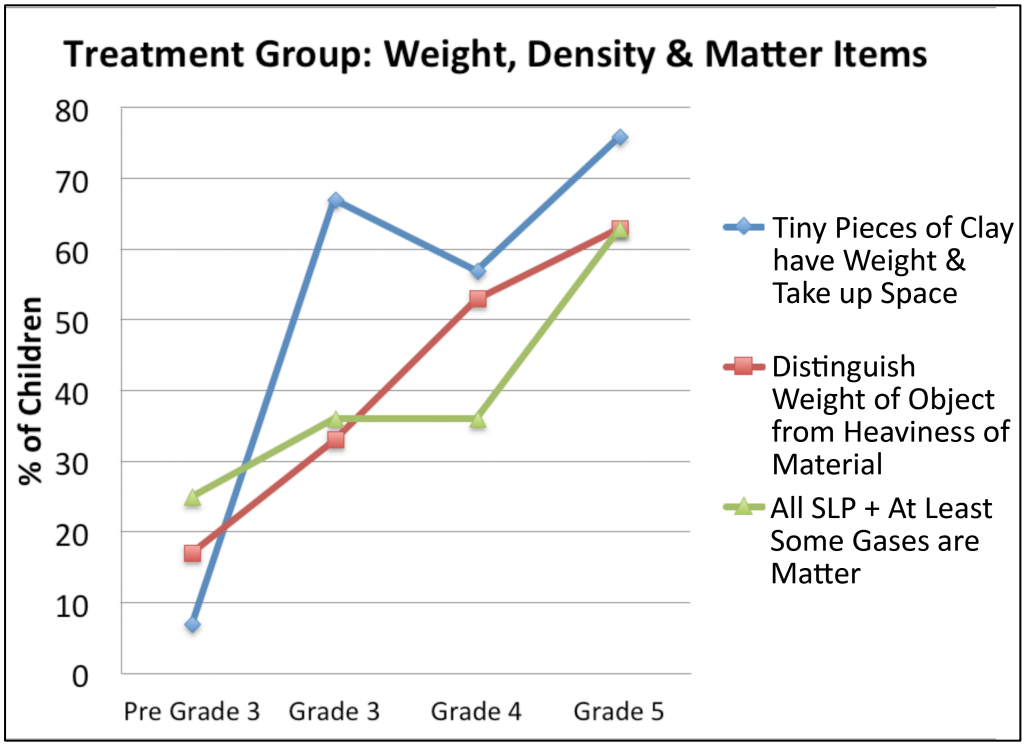

Figure 1. Grade 3 to 5 Treatment students made marked progress judging that tiny pieces take up space and have weight, distinguishing heaviness of material from weight (3 items), and including solids, liquids, gases as matter, but at different rates.

- In contrast, the progress that the Control students made in reconceptualizing weight and developing explicit concepts of density and matter were more limited (see Figure 2). Indeed, there were significant differences in the level of understanding achieved for these

core concepts by grade 5. This is not surprising given that their standard science curriculum involved more superficial and scattershot introductions to matter (e.g., separate units about weight and volume measurements and water cycle that did not connect with each other), but is consistent with the claim that more progress is made with units that focus on exploring conceptual relations in progressive fashion, and revisiting and extending topics in successive years.

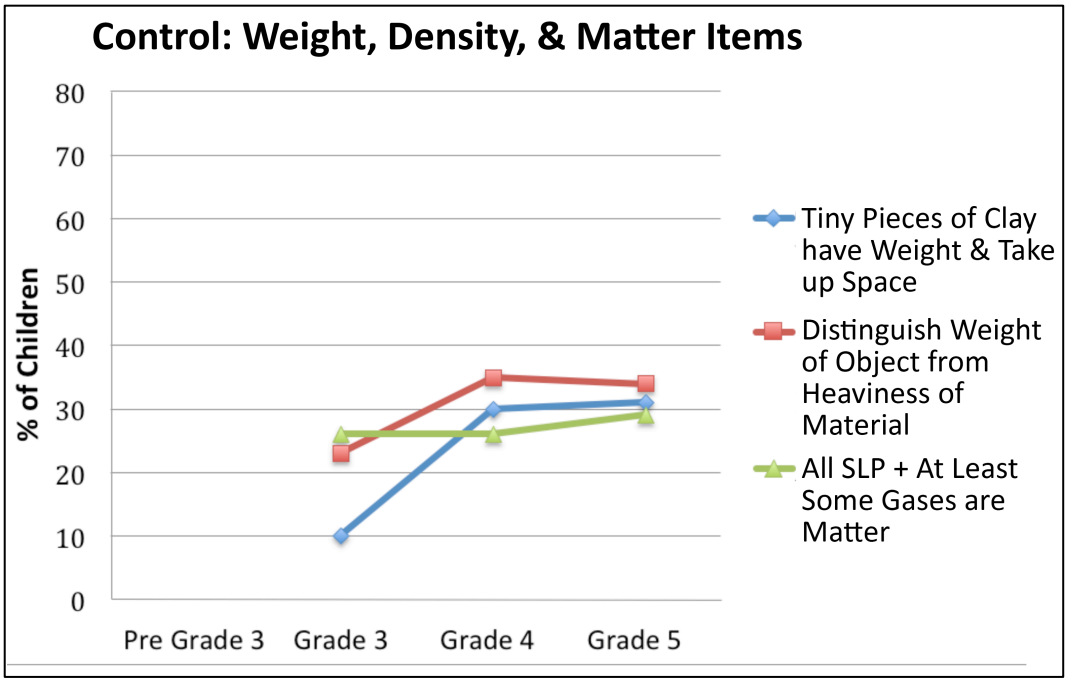

Figure 2. Grade 3 to 5 Control Students made much less progress judging that tiny pieces take up space and have weight, systematically distinguishing heaviness of material from weight, and including solids, liquids, gases as matter.

- There were, however, some tasks on which both Treatment and Control students made similar progress. For example, both Treatment and Control made significant progress in developing mathematical ideas underlying ratio and proportion (see Figure 3) – although many were far from ceiling on these important ideas by the end of 5th grade. The similar progress with these (more general mathematical ideas) may reflect that these students experienced the same math curriculum. The fact that the Inquiry students did not make more progress with these ideas may reflect on the fact that the curriculum itself did not target them sufficiently. One could imagine how a very different elementary school math curriculum (e.g., one based on ideas of modeling, such as discussed by Lehrer, Schauble, Strom, & Pligge, 2003) could work in synergy with the Inquiry curriculum would have produced even more progress with these ideas.

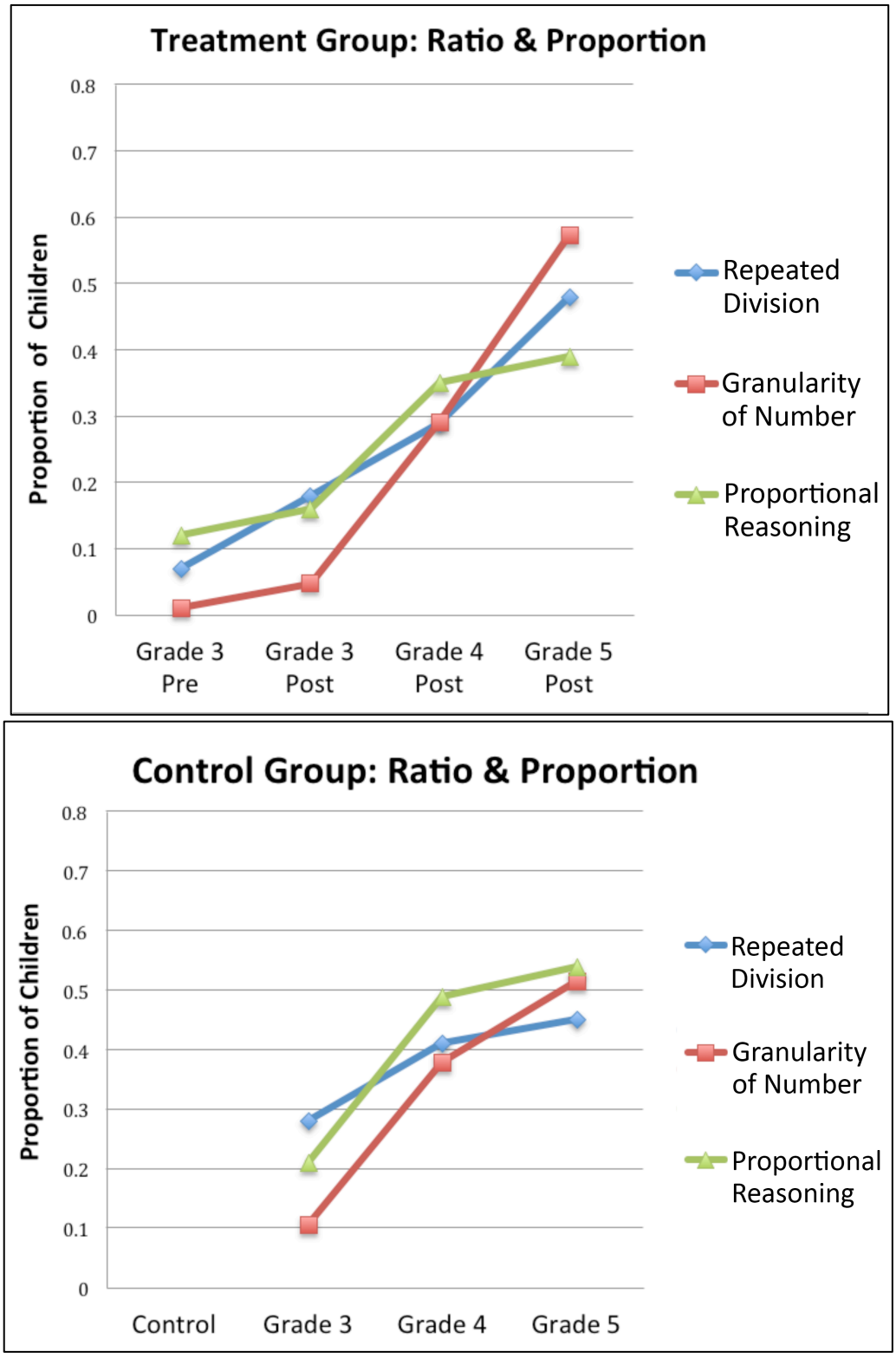

Figure 3. Treatment and Control both made mprovement in tasks probing understanding of repeated division (can keep dividing by 2 a long time or forever without getting to 0), granularity of number (there are lots or an infinite number between two integers) and proportion (correct on two items, including one where sweetness is same with both different amoutnts of sugar/water).

- Finally, although the overall growth rates are different for the two groups, many of the same patterns of relations hold among understandings, that provide clues about important conceptual dependencies. For example, tiny pieces taking up space and having weight, differentiating weight and density, and having a broad concept of matter were strongly inter-related for both Treatment and Control students. There were also strong inter-relations with matter, weight, density, and volume concepts and emerging mathematical understandings for both groups.

In developing assessments and analyzing findings, members of the Inquiry Project research group used the methods of cognitive and developmental psychologists, as this represented their areas of expertise. They did not use the tools of psychometrics, applying rigorous measurement models in analyzing their interview data. This raises three further questions of interest to this symposium and the broader LP community:

- If psychometric approaches were used, what measurement models could be used to analyze and evaluate this data?
- Can using these tools enhance our ability to test key claims? If so, how?
- What are the limitations of current measurement models for modeling this data and what challenges remain?

Currently, we (Smith, a cognitive psychologist, and Brown, a measurement expert) have begun re-analyzing this dataset using these psychometric tools. In concluding this paper, we shall present our current thinking about all three questions.

Consider first our thoughts on the first two questions -- what measurement models we are using, the nature of some of our preliminary findings, and how we think such analyses add to our ability to test important claims about LPs:

- First, we are using Rasch modeling to examine whether individual items are working as expected and show appropriate properties of a scale. Specifically, we are using a combination of Rasch's simple logistic model (Rasch, 1960/1980) for dichotomous judgments and Masters' partial credit model (Masters, 1982) for polytomous justifications. Parameters are estimated using a marginal maximum likelihood (MML) procedure (Adams et al., 1997), implemented in the psychometric software ConQuest (Wu, Adams, Wilson, & Haldane, 2007). By investigating item fit statistics (Adams & Wu, 2011; Wu, 1997), we can collect evidence of whether each item is measuring a particular aspect of conceptual understanding in a manner that is consistent with other items measuring that same aspect. An advantage of using Rasch modeling is this allows us to use *all the data* on the response to each item in relevant statistical comparisons (including partial understandings), rather than only making dichotomous comparisons of those achieving high level benchmarks of understanding (as was done in our original analysis). Similar to concerns that the dichotomous Below Proficient / Proficient metrics mandated by the No Child Left Behind Act can mask substantial growth both above and below this cutoff, Rasch analysis is expected to provide further insight about how learning unfolds before and after key transitions.
- Second, we are comparing multi-dimensional Rasch models to examine the dimensionality of our data. We have conducted preliminary analyses comparing 1D, 2D, 5D, 8D, and 9D models and found that a fully nine-dimensional model (using all 9

dimensions listed in Table 1) provides the best fit to the data, as indicated by chi-square significance testing of the improvements in total deviance. This supports our assumption that it is important to (simultaneously) track changes in multiple key ideas if one is going to understand how conceptual reorganizations occur, and stands in contrast to much of current learning progression work which traces only single dimensions of content or conceptual understanding. Of course one reason we were able to implement such a complex nine-dimensional model was that we had so many different items ($I > 90$) for each child in our 2-hour individual interview. Having this many items is rare for learning progressions work based on written tests. This meant that, in the 9D model, there are still at least 8 items per dimension, enough to provide adequate reliability of the resulting measures.

- Third, we are looking at patterns of correlation among the dimensions for the various subgroups as well. An advantage of using Rasch modeling for the correlational analysis is that scores are not in terms of simple number of items correct (on each dimension) which would only have ordinal properties, but in terms of logit values which form an interval scale. Consequently, the resulting correlations are not biased by the use of non-interval variables. Moreover, by using multi-dimensional Rasch models, we can further reduce bias in the correlations by directly modeling covariances between dimensions during estimation, rather than correlating abilities after the fact without taking into account the standard errors of the estimates. We have begun to look at these inter-correlation matrices for various subgroups, and note that, as expected, many of the dimensions are significantly inter-related, and show similar patterns of inter-relations across Treatment and Control group. At the same time interesting differences occur in patterns of correlation across grades (the dimensions typically become more highly inter-correlated as grade increases from grade 3 to grade 5) and between Treatment and Control (the Treatment students show a broader pattern of inter-correlation than the Control). This may reflect the fact that the Treatment students are achieving a more integrated and stable network of understandings – an important characteristic of a new stepping stone.
- Fourth, once the multi-dimensional space is defined, one can use it to explore different ways that individual students traverse this multi-dimensional space. We can compare the progress made by various subgroups on each of these dimensions, as a further way of testing claims that on some dimensions Treatment students made more progress than Control, while on others they made similar progress. Just because a Rasch model assumes (for a particular dimension) a consistent ordering of item difficulties does not mean that there is only one way that students can traverse a multi-dimensional space, because the abilities of individual students can vary for the different dimensions. One can ask whether there are, however, common or typical learning trajectories, which may also shed light on whether achieving some understandings serve as pre-requisites for achieving others. In particular, we plan to use the results of the psychometric analyses to conduct growth modeling. Given that we conceptualize LPs as involving change in knowledge networks that contain different types of inter-related elements, and that stepping stones involve reorganizations among existing elements along with the addition of new elements, we hope to have growth models that can capture the dynamics of growth and change in such complex knowledge networks, including the ideas of uneven growth rates, distinctions between adjustments that involve revision of existing concepts and the introduction of new ones, and the existence of "tipping" points in precipitating change.

At the same time, there are limitations and challenges for these models.

- First, Rasch models assume a simple relationship between pairs of dimensions modeled by a single covariance. This covariance is assumed to be constant across ability levels (the assumption of homoscedasticity). Yet this is unlikely to be true. Indeed from a conceptual change perspective one would expect changing relations among dimensions at different ability levels – as concepts differentiate and coalesce. And in fact, we already have some indication that this is true for our initial analyses of patterns of correlation among different dimensions (see above). Currently one of us (Brown) is developing ways to use unique construct formulations and within-item multi-dimensional psychometric models to account for heteroscedasticity, which occurs when more complex relationships exist between pairs of dimensions, such as those involving conceptual dependencies. At one extreme, such models can account for differentiations and coalescences in which dimensions appear and disappear for students of different ability.
- Second, Rasch models also assume that the relative difficulty of items remains constant across ability levels (the assumption of parameter invariance). We have evidence that, for some items, this assumption is inappropriate. For example, students with a poor understanding of matter often think (correctly) that heat is not matter, because they think that very few things are matter. Students with a moderate understanding of matter, however, tend to overextend the concept to include heat, getting this item incorrect. Finally, students with a high understanding of matter recognize that heat is not matter, getting this item correct once again. In short, the difficulty of the heat-is-matter item changes, being more difficult for advanced students than for basic students. One approach to developing Rasch scales is to exclude items such as these, as they demonstrate poor item fit, due to an item characteristic curve that is not monotonically increasing with ability. However, responses to an item like this are not random, and changes in item difficulty are predictably associated with key transitions in understanding. As such, these items contain useful information about understanding that researchers should want to incorporate into the measurement model. The saltus model (Wilson, 1989) is intended to model changes in relative item difficulty that occur as the result of key transitions in understanding, and we are currently investigating its use in this data.

In addition to having more sophisticated multi-dimensional models that can handle changing conceptual dependencies, it would be useful to have measures of degree of coherence and articulation across different dimensions. One key hypothesis in our LP work is that those who have achieved a more robust and coherent macroscopic understanding of matter would be more able to learn about atoms and molecules as constituents of matter (with understanding), avoiding the pitfalls of raisin in pudding models. If we had a better way of assessing or measuring that coherence, it would be helpful in testing this hypothesis.

Ultimately, one of the best ways to test that hypothesis is experimentally (or quasi-experimentally) with further, comparative longitudinal teaching studies – e.g., comparing the progress students make with innovative *middle school curricula* about atomic-molecular (such as IQWST) with or without LP-based matter instruction in elementary school -- to test whether those developing compositional models early in elementary school have an advantage in constructing more sophisticated particulate models (a key claim in the matter LP), or whether

most students without this prior foundation easily "catch up" with good curricula in middle school. A final challenge is such comparative longitudinal teaching studies are time consuming and difficult to do, and given family mobility, student attrition may be considerable; for this reason, it may be helpful to supplement with shorter term and more focused experimental studies of some curricular components.

References:

Adams, R., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.

Adams, R., & Wu, M. (2011). The construction and implementation of user-defined fit tests for use with marginal maximum likelihood estimation and generalised item response models. In N. J. S. Brown, B. Duckor, K. Draney & M. Wilson (Eds.), *Advances in Rasch measurement* (Vol. 2). Maple Grove, MN: JAM Press.

Amin, T., Smith, C., and Wiser, M. (2014). Student conceptions and conceptual change: Three overlapping phases of research. In N. Lederman (Ed), *Handbook of Science Education Researc*h (pp. 57-81). New York: Routledge (Taylor and Francis Publishers).

Doubler, S., Carraher, D., Tobin, R., Asbell-Clarke, J., Smith, C., & Schliemann, A. (2011). *The inquiry project: Final report.* Submitted to the National Science Foundation DRK-12 Program.

Corcoran, T., Mosher, F., & Rogat, A. (2009). *Learning progressions in science: An evidence based approach to reform.* CPRE Research Report #RR-63. Teachers College, Columbia University.

Krajcik, J., MacNeill, K., & Reiser, B. (2008). Learning goals driven design model: Developing curricular materials that align with national standards and incorporate project based pedagogy. *Science Education, 92*(1), 1–23.

Lehrer, R., Kim, M-J., Ayers, E., & Wilson, M. (in press) Towards establishing a learning progression to support the development of statistical reasoning. To appear in J. Confrey and A. Maloney (Eds.), *Learning over Time: Learning Trajectories in Mathematics Education.* Charlotte, NC: Information Age Publishers.

Lehrer, R., Schauble, L., Strom, D., & Pligge, M. (2003). Similarity of form and substance: Modeling material kind. In D. Klahr & S. Carver (Eds.), *Cognition and instruction: 25 years of progress.* (pp. 39-74) Mahwah, NJ: Erlbaum.

Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press. (Original work published in 1960).

Wilson, M. (1989). Saltus: A psychometric model of discontinuity in cognitive development. *Psychological Bulletin, 105*, 276-289.

Wiser, M., & Smith, C. (2013). Learning and teaching about matter in the middle school years: How can the atomic-molecular theory be meaningfully introduced? In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (2nd ed.), (pp. 177 – 194).New York: Routledge.

Wiser, M., Smith, C., & Doubler, S. (2012). Learning progressions as tools for curriculum development: Lessons from the Inquiry Project. In A. Alonzo & A. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (pp. 359–404). Rotterdam, the Netherlands: Sense Publishers.

Wu, M. (1997). *The development and application of a fit test for use with marginal maximum estimation and generalized item response models.* master's thesis, University of Melbourne, Victoria, Australia.

Wu, M., Adams, R., Wilson, M., & Haldane, S. (2007). ACER ConQuest version 2.0: Generalised item response modelling software [Computer software and manual]. Camberwell, Victoria, Australia: ACER Press.