ELICITING AND REPRESENTING HYBRID MENTAL MODELS

While constructing their understanding in various science areas, students go through transitional phases that may involve richly developed and consistently used mental models. These transitional models are unique cognitive structures composed of elements of both scientifically accepted and the most commonly used initial alternative models and have been previously referred to as hybrid models. In this paper we discuss the main features of Linked Item Model Analysis (LIMA) - a novel method for eliciting and representing mental models in areas where hybrid models play a role in students' learning. We developed and applied the method in the domain of sound propagation. We also present the LIMA-based assessment package for eliciting students' mental models of sound propagation, consisting of tests in different contexts and associated spreadsheet-based software, which are now available online for classroom use.

Zdeslav Hrepic, Fort Hays state University, Hays, KS 67601
Dean Zollman, Kansas State University, Manhattan, KS 66506
Sanjay Rebello, Kansas State University, Manhattan, KS 66506

Introduction

An increasing amount of research (Brown & Clement, 1992; Hrepic, Zollman, & Rebello, 2002; Otero, 2001) shows that students commonly reach a full understanding of scientific concepts only after undergoing transitional knowledge stages in which their spontaneous initially held ideas are coexistent and in various ways intertwined with scientifically accepted knowledge. This kind of knowledge has been referred to as "hybrid knowledge" (Galili, Bendall, & Goldberg, 1993).

If knowledge structures that students employ to address problems in certain domains have a certain degree of coherence and complexity we call them mental models. We adopt the understanding of the mental model as proposed by Greca and Moreira, (2002) i.e. as "an internal representation, which acts out as a structural analogue of situations or processes. Its role is to account for the individuals' reasoning both when they try to understand discourse and when they try to explain and predict the physical world behavior" (Greca & Moreira, 2002, p.108). To designate a knowledge structure as a mental model, we also require that it entails (1) a spatial configuration of identifiable kinds of things, (2) (few) principles of how the system works and (3) (certain) predictive power (diSessa, 2002).

Mental model(s) that students use define their mental model state as the pure model state (when one model is used consistently) and the mixed model state (when different models are used to answer differently, to an expert's equivalent questions) (Bao, 1999). Novices may use more than one model at the same time in order to account for contextual differences involved in the tasks at hand (Taber, 2000), which is something that experts do in some cases on purpose (e.g. apply wave and particle model for light in different situations).

Another model state is the hybrid model state (Hrepic, 2002; Hrepic et al., 2002) in which only one so- called hybrid model is used. A hybrid model is a composite mental model that systematically combines different features of two other parental models. We allow that parental models may or may not be students' common initial alternative model and the scientifically accepted model. However that is what they typically are (Hrepic et al., 2002; Vosniadou, 1994). We also require that a hybrid model is inconsistent (by one or more features) with both models from which it is derived (Hrepic, 2002; Hrepic et al., 2002). According to this definition, a hybrid model does not include usage of different models in different instances related to the same topic (mixed model state) as found in Vosniadou's (1994) analysis who first proposed a definition of what she called "synthetic model." Greca and Moreira (2002) used the term hybrid model to denote the outcomes of recursive reformulations of students' initial models and they considered these models of the same kind as those described by Vosniadou (1994).

Hybrid mental models have been relatively recently identified in various physics and science topics ranging from earth science (Vosniadou, 1994), to electrostatics (Otero, 2001), Newtonian mechanics, (Hrepic, 2002; Itza-Ortiz, Rebello, & Zollman, 2004) and sound (Hrepic, 2002; Hrepic et al., 2002). What Galili *et al*. (1993) refer to as "hybrid knowledge" in optics seems to be fitting our notion of a hybrid model as well. This is the also the case with Brown and Clement's (1992) notion of "intermediate concepts" identified in domains of inertia and gravity. This shows that hybrid mental models have to be taken seriously while addressing students' understanding of many major science topics. It is also likely that hybrid models as transitional cognitive elements exist in other domains where they were not yet identified or described as such. Another reason to take hybrid models seriously is because while using a hybrid, i.e. an incorrect model, a student can give correct answers (false positives) to a variety of standard questions and even achieve high scores on tests, which will fail to diagnose hybrid models (Hrepic, 2002).

In this study we created a formative assessment of mental models of sound propagation and employed a novel testing and analytical method to address and represent students' hybrid models.

Goals and Research Questions

The goal of this study was to develop a multiple-choice test that can elicit students' mental models of sound propagation during the lecture while using a classroom response system and appropriate software. The difficulty with this task is that mental models may not be, and frequently are not stable, especially with novices. They may be incomplete and may contain contradictory elements (Norman, 1983; Redish, 1994). However, when the learning of a particular physics topic is explored through systematic qualitative research, usually a small finite set of commonly recognized models is identified (Marton, 1986).

This finding is a basis for quantitative approaches to eliciting mental models such as Model Analysis (Bao & Redish, 2001; Bao, Zollman, Hogg, & Redish, 2000). This approach "assumes that the most commonly used mental models are identified through extensive qualitative research. These known models can then be mapped onto the choices of an appropriately designed multiple-choice test" (Bao & Redish, 2001, p.3).

Our approach to quantitative analysis of mental models starts with these same assumptions. This is possible because previous research related to students understanding of sound (Hrepic, 2002, and the references therein) consistently show that there is a small number of fundamental ideas that students express about nature of sound propagation. At the most essential level, these ideas boil down to wave-like and object-like model of propagation (e.g. Barman, Barman, & Miller, 1996; Hrepic, 2002; e.g. Linder & Erickson, 1989; Maurines, 1993; Wittmann, 2001).

In our previous research (Hrepic, 2002; Hrepic et al., 2002) we investigated students' mental models of sound propagation through in-depth interviews of 23 students, 16 of which were interviewed both before and after instruction. The students were enrolled in a conceptual-level introductory physics course at Kansas State University (KSU). The study showed that most of the students (78% of students in 69% of the interviews) express their ideas related to sound propagation in a way consistent with our definition of a mental model. In remaining instances the model was not clearly expressed, but each student's answers on conceptual questions were consistent with either one or two models that other students described (but not fully or indisputably expressed). Additionally, the study showed that even when students are asked open ended questions in an interview setting, most of them describe only a small set of commonly shared mental models. These findings give ground to consider elicitation of mental models of sound propagation on a large scale a meaningful and potentially, instructionally productive endeavor.

Our findings (Hrepic et al., 2002) build on and are consistent with other, earlier research on students understanding of sound propagation mentioned before. Based on these studies, the following four fundamental models of sound propagation can be distinguished: Wave Model (scientifically accepted); Intrinsic Model (Hybrid); Dependent Entity Model (Hybrid) and Independent Entity Model (common initial alternative model). The fundamental model in this context refers to a model with a set of features that one or more students' mental models identified through the qualitative research have in common.

According to the first two models sound is a specific vibrational (Wave Model), i.e. translational (Intrinsic Model) motion of particles of the medium caused by the source of sound. According to Entity Models, sound is a self-standing entity different from the medium through which it propagates. In order to propagate, sound needs (Dependent Entity), i.e. does not need (Independent Entity) the particles of the medium and their motion. So, according to the Independent Entity Model sound can propagate through the vacuum. Each of these models also has a range of variations or sub-models that our test probes for (e.g. wave model can be longitudinal, transversal and circular).

In addition to the models of sound propagation, there is a specific understanding of what the sound is that may be associated with different mechanisms of propagation. This understanding is that the sound is what we hear, i.e. it is exclusively what we hear, and we refer to this understanding as “the Ear-born Model.” A feature of the Ear-born sound is that it is a partially correct idea and is well aligned with our everyday definition of the sound.

All of the fundamental models described above (as well as their variations or sub-models) differ one from another according to the answers that they give for the four questions. (a) What is sound? (b) What happens to the sound without the medium? (c) What are the dynamics of the particles of the medium during the sound propagation? (d) How are these dynamics related to the sound propagation? Our task was to create a test that can elicit these models and their sub-models in a classroom situation in real time.

Research questions

Research questions that we address in this paper are:

1. Are there any other ideas about nature of sound propagation that do not fall within the bounds of students’ common fundamental models of sound propagation elicited in previous studies?
2. How and to what extent can we elicit mental models of sound propagation through a multiple choice instrument?
3. How do we represent the data on students’ models so that the representation provides a variety of information to guide instruction even in real time?

We approached these research questions keeping in mind limitations imposed by the unstable nature of (particularly novices’) mental models. Also, although mental models of sound propagation that we want to probe have been elicited through extensive and careful qualitative and semi-qualitative studies, they are nevertheless our i.e. the researcher's models about what students are thinking.

Like many other things that scientists (e.g. physicists) try to describe (such as subatomic particles), we may never be able to see or ‘read’ what's in a student's mind, but we can (like scientists often do) construct a model (based on experimental evidence) about what or how a student might be thinking (based on what they tell us). In this sense, the term “mental model” that we want to elicit applies to our model about the students' model. The reasons to build models of student thinking are similar to reasons for which physicists build models: such models provide us with a vocabulary or framework to describe the topic at hand, which is in our case - student’s knowledge and difficulties s/he has. Models also, when carefully measured and described, have predictive power. Therefore they can enable the instructor to predict students' performance and can provide the useful feedback to remedy students’ problems. This research has created an instrument that can in real time determine what model best describes a student’s thinking as well as understanding of a class as a whole.

Methodology

To address our research questions we employed both quantitative and qualitative methods. Our underlying theoretical paradigm is constructivism although in some aspects our approach toward elicitation of mental models may have elements of positivism. We divided the research procedure into four major steps that we labeled pilot testing, pre-survey testing, survey testing and post-survey testing. Each of these steps is described in the following sections.

Pilot Testing

Pilot testing had two purposes. The first one was to determine if anything of significance was omitted in earlier qualitative research on a relatively small sample of 23 students in terms of students' ideas related to sound propagation. To answer this question we administered an open-ended questionnaire, similar to our interview protocol, to another 158 students enrolled in large concept-based introductory physics class at KSU. The second purpose of pilot testing was to determine the optimal contextual situations for eliciting students' models of sound propagation. To address this question, researchers administered a battery of semi-structured conceptual questions related to sound both in general and in a variety of specific situations. The survey was administered before and after instruction to another large enrollment arithmetic-based introductory physics class at KSU. Out of 139 students enrolled in the class, 128 took the pre-instruction test and 115 took the post-instruction test.

Pre-Survey Testing

Once the models we want to elicit were known, along with the optimal contextual situations, the next step in the test creation was mapping the defined mental models onto the answer choices of the multiple-choice test. In the pre-survey phase, the first version of the multiple-choice test was probed and then refined.

For this purpose we initially utilized a survey in which, in addition to the model-related choices, each question had the option to list more than one choice or to write in an independent answer different from any of the choices provided. These additional two choices were included to determine the need for possible adjustments of the offered choices and to determine the possible need to include new choices. During this stage of research we interchangeably used interviews and surveys at a larger scale (N>30 each time) and were changing the test as we gathered more data and feedback from the students. Students who participated in the study at this stage were enrolled in all three levels of introductory courses at KSU (concept-based, algebra-based and calculus-based). In addition, at this stage of the research we made an initial validation of the test through experts' review of the test.

Survey Testing

After the detailed preparation of the test questions, answer choices and their wording, we administered the test to a large number of students (1600 at this stage) at 13 different educational institutions in the US and Croatia. The Croatian language version was translated from English by the first author who is a native Croatian speaker and who is also proficient in English. It was than validated by a certified court interpreter for English language in Croatia. The purpose of this survey administration was to determine whether the answer choices are correlated in a meaningful way when data is collected from a large number of students in a variety of institutions and educational settings.

We sent out requests for participation in research to physics instructors in the US through the most active physics education research e-mail list (PhysLrnr) in the US and possibly internationally as well. In addition we sent the same request to physics instructors that the author knew in his native country of Croatia and several of them administered the test in their classes. Therefore, the test was administered in all classes whose instructors agreed to participate in the study regardless of the character and level of the institution or the class size and its instructional setting. In this sense the sampling procedure was neither random nor one of convenience.

We combined this quantitative analysis with validation by interviewing KSU students enrolled at three different levels of introductory physics courses and comparing their open ended answers with choices they selected on the test.

Post-Survey Testing

Based on the findings during the survey part of the study we finalized the test by making adjustments in few of the answer choices and by adding a figure describing the situation with a bell in a vacuum. The new version was than validated once again through correlation analysis and expert validation. Finally, we performed a role-playing validation in which participants with a Ph.D. degree in physics played the role of students "having" different mental models of sound propagation. Based on "their" models, participants were supposed to pick the answers in the test that corresponded to their models.

Findings and discussion

In this section we describe findings as they pertain to research questions. The test validity and reliability is addressed separately in the Appendix A.

The first research question was addressed in the pilot phase of the research. All ideas that were expressed in the open-ended answers are consistent with mechanisms of sound propagation that were identified earlier and no new or different ideas were found. Although we cannot exclude the possibility that a student in some other population may have an idea of propagation that does not fit any of fundamental models that we identified, these instances, if they occur, will be rare and isolated. Another implication of the pilot testing was that the contextual situations optimal for elicitation of mental models of sound propagation are simple propagation in the air, propagation through a barrier (wall) and "propagation" through the vacuum.

Issues in Eliciting Hybrid Models

The main difficulty in mapping defined mental models onto answer choices of the multiple-choice test is that these mental models cannot be reduced to simple knowledge elements. Rather, they are stories that may not fit a single answer choice. Because a model can have sub-models, more than one answer choice in a question may correspond to the same model. Also, because of the nature of hybrid models, more than one model may be associated with the same choice. For example, longitudinal movement of medium particles during sound propagation is consistent with the Wave, Independent Entity, Dependent Entity and Ear-born Models. Hybrid Models also cause overlaps in multiple-choice answers so that in some cases different choices pertaining to the same question may have substantial commonalities. For example, according to both the Independent and Dependent Entity Models, if sound is created in the medium it passes through empty spaces in between the particles of the medium. While constructing the test we also wanted to avoid situations in which a student gives correct answers to a question while using a hybrid model. Through several iterations during the pre-survey phase of research, we developed and implemented an inventory that addresses the aforementioned issues. In the process we combined data from interviews with students, correlation analyses of answer choices obtained from students enrolled in different introductory physics courses at KSU and validation and inputs from experts.

Sound Model Inventory

The result of this study is a model inventory that elicits mental models of sound propagation in real time, i.e. the inventory can be utilized during the instruction as a formative assessment. The full name of the test is “Formative Assessment of Mental Models of Sound Propagation” or for short “FAMM-Sound.”

A minimum of three different questions are needed to probe a student’s model. Two questions (Q2 & Q3) were needed to elicit the movements of the particles of the medium related to the sound (in order to allow for a variety of movements that students express in open-ended questions). One more question was necessary to associate this motion to sound propagation. Additional test questions were used to determine students’ consistency, i.e. their model state. There are two tests: one for propagation through air (and vacuum) and through a wall (and vacuum). Below we have paraphrased the test questions for both contexts:

1. What is ***basic mechanism of sound propagation*** in the air/wall?
2. How do particles of the medium ***vibrate***, if at all, while the sound propagates?
3. How do particles of the medium ***travel***, if at all, while the sound propagates?
4. What does this motion have to do with sound propagation – ***cause and effect relationship***?
5. What does this motion have to do with sound propagation – ***time relationship***?
6. What happens with sound ***propagation in the vacuum***?

Linked Item Model Analysis (LIMA)

Due to aforementioned problems with eliciting hybrid models, it was not possible in this test to map different models onto answer choices so that each choice corresponds to a unique model. This restriction was the primary reason that the Model Analysis method as described by Bao (2000) was not useful in this case. Instead, the Linked Item Model Analysis approach was developed and utilized in a following way:

The test results are analyzed to first determine if the student is self-consistent, i.e. if he or she uses a single model throughout the test. A program compares student's set of six answers with sets of answer combinations associated with the models. If a match is found, the student is consistent or in a pure (un-mixed) model state (which may or may not be associated with a hybrid model). If no match is found, the student is in a mixed model state (inconsistent). In that case, the analysis program identifies the different models that the student uses by combining answers from questions Q2 and Q3 (dynamics defining questions) with each of the remaining four questions respectively. This way each test (air context test and wall context test) probes a student's model four times (through these four question-triplets). In the air context, a minimum of three questions are needed to determine the model once. In the wall context, which is more complex than the air context because of the larger number of the factors involved, four and sometimes all six test questions are used to determine the model associated with a particular answer choice. A model is not necessarily ascribed to any triplet and student may be classified into the "other" category associated with no-model or un-identified model.

Because the situation in the wall-vacuum context involved one additional factor (wall particles) with respect to the air-vacuum context, sometimes four or even all six questions were needed in order to determine the meaning of a single answer choice in a certain question.

We call this new approach to model analysis in which the model associated with a particular answer choice is determined by answers given in other (sometimes all other) test questions the Linked Item Model Analysis (LIMA). This approach makes it possible to address all of the issues regarding hybrid models mentioned above and to elicit students' mental models (including hybrid models) and their model states (including hybrid model state).

Display of Results in Terms of Mental Models

The Excel®-based analysis program displays results in the five different graphs that show: (1) Percentages of times that a particular model is used by a class as a whole (see Figure 1.), (2) Percentages of students using a particular model at least once, (3) Movements of particles of the medium, (4) Students' model states, and (5) correctness of the answers. Figure 1 shows the graph that displays percentages of times that a particular model is used (with respect to possible number of times that the model could have been used). Model usage displayed in Figure 1 shows all fundamental models separately as well as the Ear-born Model so that contributions from the consistent and inconsistent usage of each of them are stacked in the same column.

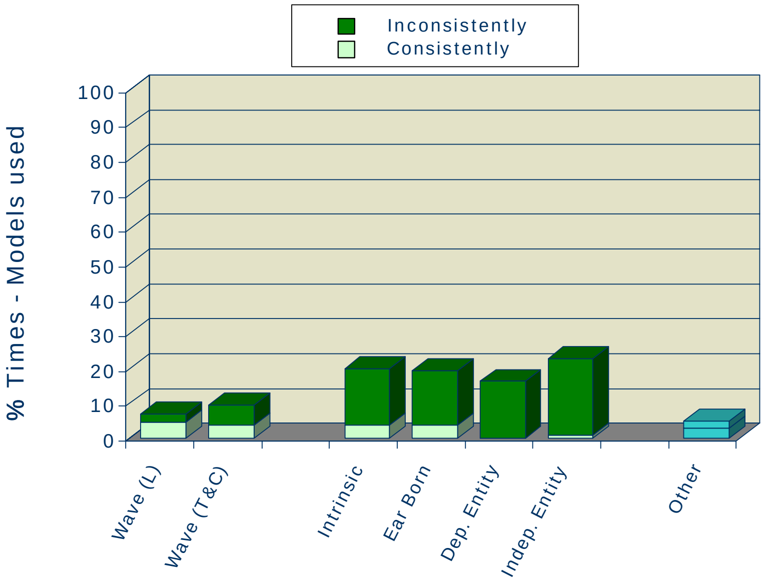

Figure 1. Percentages of times each model is used.

Movements of particles of the medium are displayed in a different bar chart (Figure 2). Horizontal axes display different vibrations and translational motion is added on top of each of them according to the combinations that students picked.

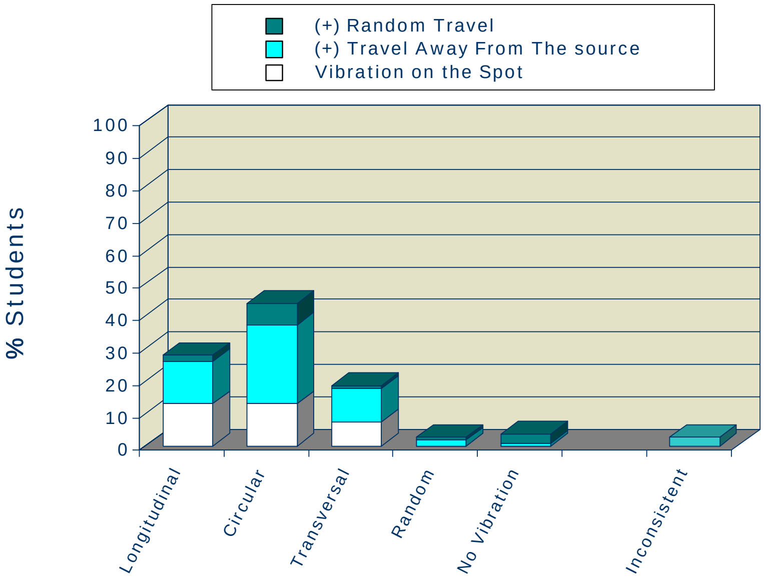

Figure 2. Movements of particles of the medium.

We show one of the remaining three graphs in Figure 3 which displays students' model states. The two bars show the number of students in the pure and mixed model states. Each bar is parsed to provide additional information to guide instruction. The pure model state column displays students that consistently use one model and it separates students that use Wave (correct) model and those that use any of the incorrect models. In the column displaying mixed model states students that mix only particular combinations of models are separated from others. Students that mix only Wave and Ear-born Models are separated because one might argue that there is nothing wrong with this combination if put together. A second distinguished mixture of models is the combination of Independent Entity and Dependent Entity Models. These two models are not separated by a clearcut borderline but lay along the continuum of "dependency." For example, the statement that the medium "helps sound propagation" (but it's not necessary) would lay somewhere in the middle of this "dependency" continuum. This continuity together with combination of contexts employed in each of the tests (propagation through the medium and through the vacuum) causes relatively frequent mixtures of exclusively Independent and Dependent Entity models.

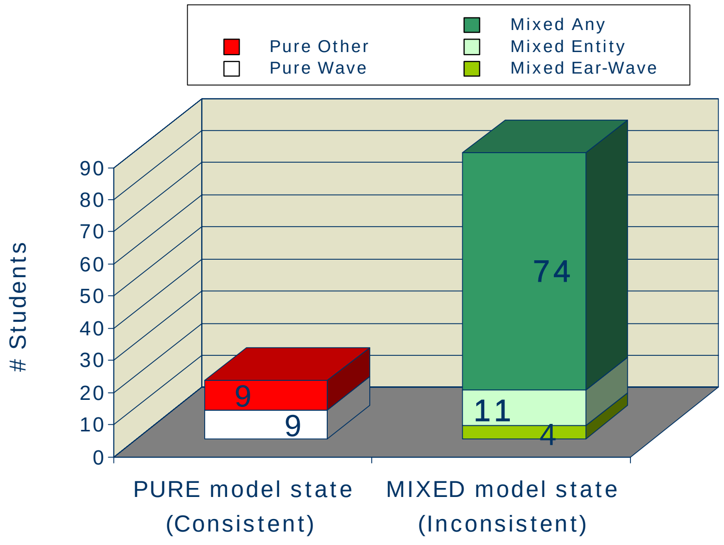

Figure 3. Students' model states.

Results in sample graphs shown in Figures 1-3 are from a university algebra-based introductory physics course (2003, post-instruction survey). In addition to five graphs (three of which were discussed above), the analysis program contains a sheet with a detailed analysis of student use of each of the sub-models (model variations).

<u>Range of unclassifiable model probes</u>

Each possible answer combination does not necessarily correspond to a model. We hoped that the percentage of combinations that do not correspond to an identifiable model will be less than 10%. At the university level we found that the average number of such combinations (not identified to a model and classified as "other") in all the tests taken during the survey phase was 5.21% (N_{tests}=1281). At High school level the average percentage of "others" was 9.78% (N_{tests}=300) and at middle school level it was 8.55% (N_{tests}=132). This makes the percentage of unclassified answer combinations less than 10% at any educational level. In another words the test elicits students' models in more than 90% of the cases which makes it a useful formative instrument.

Test Results, Validity and Reliability

In the Appendix A, we show results obtained during the survey and post-survey phases of the research. In the Appendix B we built the case that the test is a valid instrument based on survey results and validity-verifying procedures employed in the study. We present a range of arguments to demonstrate that this test is an assessment tool that reliably and validly addresses students' mental models of sound propagation at the high school and college level. This is especially applicable when the test is used as a formative assessment tool because test validity is not an attribute of the test, but "of the interaction of a test with a situation in which the test is used to make decisions" (Hanna, 1993, p. 382). Based on this test, a decision about optimal instructional approach related to sound propagation can be confidently made.

The FAMM-Sound test and associated analytical software, together with the dissertation (Hrepic, 2004) demonstrating further details related to the test's validity and reliability can be downloaded for personal use from: http://web.phys.ksu.edu/role/sound/

Applicability of the test at different levels

Results obtained at the high school and college level that pertain to reliability of the test are similar and both are very good (see Appendices A and B). Validity of the test was established for the college level through interviews with the students in addition to correlation analysis of the survey results. At the high school level validity was established through quantitative (analysis of survey results) but not qualitative procedures (interviews). We also established test validity through a series of generic procedures (such as expert validation, role playing etc., which are described in the Appendix B). The implication is that we can claim the test is a valid instrument at both college and high school levels but the case that we made for its validity is stronger case for the college level. The test also shows promising results at the middle school level, but results collected so far are inconclusive and it is not clear whether students at this level validly interpret the test items (see Appendix B).

Test limitations

Several limitations of the test became evident throughout the test validation process and these are, to a large extent, unavoidable with any multiple-choice instrument. (1) The test affects students' understanding in that test items (questions and answer choices) play a significant role in the change and dynamics of students'model (2) Students' test-taking strategies, that are otherwise meaningful and effective may obscure test results. (3) The test may identify a "no model" state as a mixed model state and possibly even as a pure model state if a student picks a model consistently. (4) Students may change their opinion as they take the test, without being aware of this change.

Although these limitations are typical for the multiple-choice tests, they deserve to be mentioned because the user should have them in mind when interpreting the test results.

Conclusions and Suggestions for Further Studies

In summary, based on this and previous research, we conclude that common fundamental models of sound propagation (that we described and classified into broad categories of Wave, Intrinsic, Dependent Entity, and Independent Entity Models) adequately describe all of the identified students' ideas related to sound propagation.

These models can not be elicited through multiple-choice questionnaires in a way that one of model is mapped on one answer choice. Rather answer choices in several questions have to be combined for this purpose. We developed and employed this method of model eliciting and called it LIMA (Linked Item Model Analysis). Our results show that the test to that we developed together with his analysis of method elicits students' mental models of sound propagation in a more than 90% of instances at the post-secondary, secondary and middle school levels. However, validity of the test at the middle school level should be further investigated. We represent data on students' models through five different graphs which provide a variety of instruction guiding information and can be employed in real time.

The difference between the analytical method of analysis of students' model states developed in this study and those suggested earlier (Bao, 1999), is that in this approach there is no one-to-one match between answer options and mental models. Another major difference is in the representation of the results in terms of students' usage of mental models. Our aim was to graphically represent students' usage of each of the models (including hybrids) separately as well as the students' consistency so that all of this information is available and easy to read while the test is used as a real-time classroom formative assessment. Finally, our approach does not treat students' model states probabilistically although we are aware of our current limitations in understanding the complexity of mental processes involved in conceptual change.

Suggestions for further research on this topic include investigating (1) whether LIMA can also be useful in topics where hybrid models are not necessarily dominant, (2) whether this approach to addressing models in real time can facilitate the desired conceptual change, (3) how effectively teachers can implement real-time formative evaluation using this testing approach, (4) whether this testing approach is applicable for eliciting other psychological constructs (not necessarily cognitive ones) such as personality tests and (5) whether this test provides useful information that currently available tests in that field do not? Namely, it is possible that personality tests in which answers on different questions are combined into full sentences might provide insights into the examinees' psychological states that are missed in inventories with self-standing questions. If this is the case, LIMA might find its applications not only in science education but in psychology as well.

Acknowledgements

Work supported in part by NSF grant # REC-0087788.

References

Bao, L. (1999). *Dynamics of student modeling: A theory, algorithms, and application to quantum mechanics.* Unpublished Ph.D. Disertation, University of Maryland, College Park, MD.

Bao, L., & Redish, E. F. (2001). Model analysis: Assessing the dynamics of student learning. *Submitted to Cognition & Instruction*.

Bao, L., Zollman, D., Hogg, K., & Redish, E. F. (2000). Model analysis of fine structures of student models: An example with newton's third law. *Journal of Physics Education Research, submitted*.

Barman, C. R., Barman, N. S., & Miller, J. A. (1996). Two teaching methods and students' understanding of sound. *School Science and Mathematics, 2,* 63-67.

Brown, D., & Clement, J. (1992). Clasroom teaching experiments in mechanics. In R. Duit, Goldberg, F., Niedderer, H. (Ed.), *Research in physics learning: Theoretical issues and empirical studies* (pp. 380-389). Kiel: IPN.

diSessa, A. A. (2002). Why "conceptual ecology" is a good idea. In M. Limon & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 29-60). Dordrecht, Netherlands: Kluwer Academic Publishers.

Galili, I., Bendall, S., & Goldberg, F. M. (1993). The effects of prior knowledge and instruction on understanding image formation. *Journal of Research in Science Teaching, 30*(3), 271-301.

Greca, I. M., & Moreira, M. A. (2002). Mental, physical, and mathematical models in the teaching and learning of physics. *Science Education*, *86*(1), 106-121.

Hake, R. R. (1997). Interactive-engagement vs traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *Am. J. Phys, 66*, 64-74.

Hanna, G. S. (1993). *Better teaching through better measurement*.Orlando, Florida: Harcourt Brace Jovanovic, Inc.

Hrepic, Z. (2002). *Identifying students' mental models of sound propagation.* Unpublished Master's thesis, Kansas State University, Manhattan, KS.

Hrepic, Z. (2004). *Development of a real-time assessment of students' mental models of sound propagation.* Unpublished Doctoral dissertation, Kansas State University, Manhattan, KS.

Hrepic, Z., Zollman, D., & Rebello, S. (2002). *Identifying students' models of sound propagation.* Paper presented at the 2002 Physics Education Research Conference, Boise ID.

Itza-Ortiz, S. F., Rebello, S., & Zollman, D. A. (2004). Students' models of newton's second law in mechanics and electromagnetism. *European Journal of Physics, 25,* 81–89.

Linder, C. J., & Erickson, G. L. (1989). A study of tertiary physics students' conceptualizations of sound. *International Journal of Science Education, 11*(special issue), 491-501.

Marton, F. (1986). Phenomenography: A research approach to investigating different understandings of reality. *Journal of Thought, 21*(3), 28-49.

Maurines, L. (1993). Spontaneous reasoning on the propagation of sound. In J. Novak (Ed.), *Proceedings of the third international seminar on misconceptions and* *educational strategies in science and mathematics*.Ithaca, New York: Cornell University (distributed electronically).

Norman, D. A. (1983). Some observations on mental models. In D. A. Gentner & A. L. Stevens (Eds.), *Mental models*.Hillsdale, NJ: Lawrence Erlbaum.

Otero, V. K. (2001). *The process of learning about static electricity and the role of the computer simulator*. Unpublished Ph.D. Dissertation, University of California, San Diego, CA.

Redish, E. F. (1994). The implications of cognitive studies for teaching physics. *American Journal of Physics, 62*(6), 796-803.

Taber, K. S. (2000). Multiple frameworks? Evidence of manifold conceptions in individual cognitive structure. *International Journal of Science Education*, *22*(4), 399-417.

Vosniadou, S. (1994). Capturing and modeling the process of conceptual change. *Learning & Instruction*, *4*, 45-69.

Wittmann, M. C. (2001). The object coordination class applied to wavepulses: Analysing student reasoning in wave physcs. *International Journal of Science Education, 24*(1), 97-118.

Appendix A: Test Results Obtained In the Survey Phase

The results of the testing in the survey phase are summarized in Table 1 in terms of the model distribution and students' self consistency. The results are presented separately for pre and post instruction results obtained from students at different educational levels. The percentages in Table 1 reflect simple averages of the percentages that each of the models was used in each of the samples pertaining to the specific category. A standard deviation was calculated with respect to these simple averages.

While calculating the averages we excluded samples that had fewer than 15 students. We also excluded samples that were tested after instruction during which a particular intervention was made in order to address students' understanding of the sound that would not have been made without the test or without the instructors' familiarity with the test. These samples were excluded because they were not compatible with others and can be considered outliers by their characteristics although they were not necessarily outliers by results. The number of samples that were included and the number of incompatible samples in each of the analyzed categories is reported in Table 1.

Table 1.

Results of the surveys in terms of the model distribution and students' self consistency (in percentages).

Context	Pre / Post	Institution	Included Samples	Incompatible samples	N	Result	Consistent (Pure model state)	Consistent Wave	Wave (L)	Wave (T&C)	*Intrinsic*	Ear-born	Dependent Entity	Independent Entity	Other
Air	**Pre**	University	3	1	257	Average SD	**14.05** 8.16	**4.26** 4.11	**3.99** 4.84	**5.41** 2.75	**22.93** 2.97	**16.64** 4.92	**18.28** 1.67	**27.84** 6.64	**4.24** 1.46
Air	**Pre**	High School	1	0	28	Average	**7.14**	**0.00**	**0.89**	**5.36**	**13.39**	**14.29**	**28.57**	**26.79**	**10.71**
Air	**Post**	University, CC	11, 1	3	689	Average SD	**13.60** 11.36	**5.66** 6.89	**7.28** 8.24	**4.35** 3.76	**21.13** 6.91	**13.78** 5.84	**19.67** 7.50	**29.00** 7.49	**4.80** 3.89
Air	**Post**	High School	3	2	156	Average SD	**14.67** 6.65	**1.71** 2.01	**4.71** 3.65	**2.08** 2.22	**15.97** 4.14	**16.29** 3.88	**20.14** 2.19	**32.64** 2.01	**8.16** 1.43
Air	**Post**	Middle School	2	0	64	Average SD	**11.14** 12.54	**0.00** 0.00	**0.00** 0.00	**0.00** 0.00	**13.75** 1.77	**36.99** 11.33	**22.27** 1.45	**20.23** 9.16	**6.76** 2.49
Wall	**Pre**	University	1	1	76	Average	**14.47**	**3.95**	**3.29**	**6.91**	**12.83**	**6.25**	**23.03**	**43.09**	**4.61**
Wall	**Pre**	High School	1	1	21	Average	**14.29**	**0.00**	**0.00**	**1.19**	**23.81**	**7.14**	**21.43**	**26.19**	**20.24**
Wall	**Post**	University, CC	6, 1	3	338	Average SD	**13.45** 11.49	**5.56** 7.70	**6.20** 7.20	**8.06** 5.06	**19.01** 8.14	**2.90** 3.30	**24.37** 6.58	**32.66** 13.32	**6.78** 5.26
Wall	**Post**	High School	3	2	95	Average SD	**15.05** 2.57	**5.75** 2.30	**7.90** 4.32	**6.48** 2.41	**14.76** 2.38	**8.19** 1.88	**22.77** 2.69	**30.02** 0.86	**9.87** 7.61
Wall	**Post**	Middle School	2	0	68	Average SD	**7.81** 1.82	**4.55** 6.43	**0.27** 0.38	**10.20** 12.89	**14.23** 3.98	**4.10** 4.19	**31.18** 4.12	**29.77** 14.78	**10.25** 1.64
Both	**Pre**	University	1	1	175	Average	**13.14**	**2.86**	**2.57**	**4.43**	**17.86**	**13.86**	**20.00**	**37.14**	**3.57**
Both	**Pre**	High School	1	0	49	Average	**10.20**	**0.00**	**0.51**	**3.57**	**17.86**	**11.22**	**25.51**	**26.53**	**14.80**
Both	**Post**	University, CC	6, 1	2	559	Average SD	**10.93** 9.34	**4.27** 5.31	**5.71** 5.13	**5.72** 4.09	**21.07** 7.29	**8.31** 3.90	**21.62** 7.10	**31.09** 9.77	**6.47** 4.66
Both	**Post**	High School	3	2	251	Average SD	**14.17** 3.84	**3.38** 1.24	**5.19** 2.42	**4.11** 1.84	**15.65** 3.73	**12.75** 2.75	**21.05** 0.67	**31.39** 0.67	**9.85** 1.96
Both	**Post**	Middle School	2	0	132	Average SD	**9.37** 6.96	**2.38** 3.37	**0.14** 0.20	**5.34** 6.76	**14.01** 2.92	**19.90** 3.00	**26.90** 1.52	**25.14** 11.98	**8.57** 2.02

Table 1 shows that the obtained results are stable in several different ways: across different educational levels, across different course levels at the same institution, across the same levels at different institutions and between pre- and post-instruction tests. These results are important for determining the test reliability so we elaborate them in the sections below.

Stability of the results across different educational levels

Table 1 show that differences between different educational levels are in the expected direction. Namely, college students perform better than high school students and high school students perform better than middle school students. This is the case with correctness of the answers as well as students' self-consistency. An exception, however, is in the case of the post-instruction tests and the wall context. In surveyed samples, high school students on average outperformed the college students. However, this is not the case with the air context alone nor is it the case when two contexts (all students) are taken together. It should be noted, however, that all of these differences between educational levels (in terms of students' self-consistency and in terms of usage of the correct model) are embarrassingly small for higher levels with respect to the lower ones.

Figure 4 graphically compares post-instruction results at three educational levels as obtained through the air context of the test. The figure shows stable increasing slope when Wave and Intrinsic Models are compared at different levels. The Ear-born Model has the opposite trend, which (much less pronounced) exists also in the case of the Dependent Entity Model. There is no real pattern of this kind in the case of the Independent Entity Model.

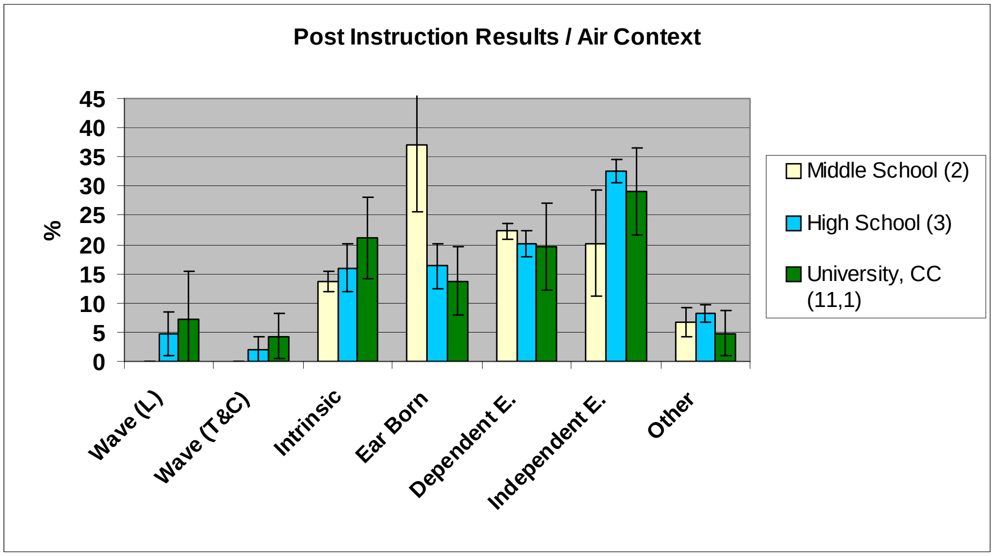

Figure 4. Comparison of post instruction results at primary, secondary and tertiary levels as obtained by the air context of the survey.

The differences shown in Figure 4 are easier to notice if models that are similar to some extent are grouped together. In this way we can group Wave Models and Intrinsic

Models because the same answer choices correspond to these models in questions 1, 4, 5 and 6 and they are differentiated by the dynamics of the particles of the medium in questions 2 and 3. Dependent and Independent models have in common that according to both of them sound is a self-standing entity different from the medium through which it propagates. If these two groups of models are clustered together, Figure 4 appears as shown in Figure 5.

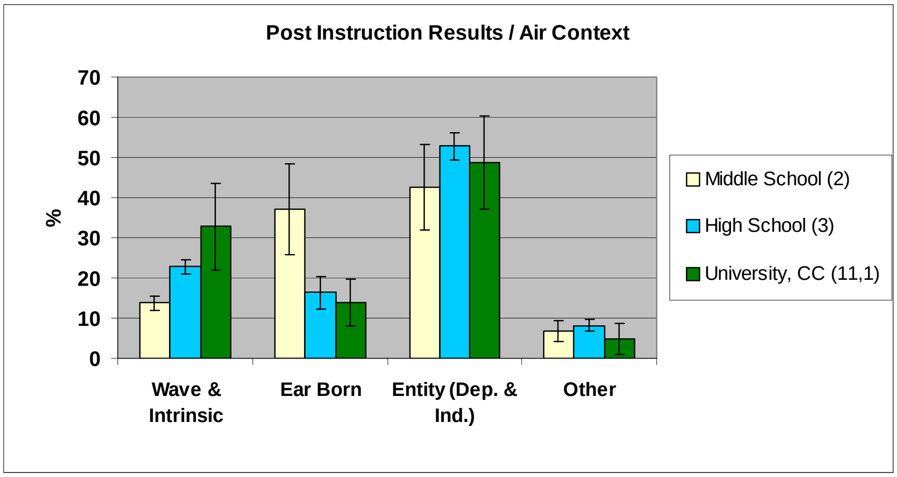

Figure 5. Comparison of post-instruction results at primary, secondary and tertiary levels as obtained by the air context of the survey (grouped models).

When models are grouped this way, patterns described with respect to Figure 4 become more pronounced. There is an upward slope when Wave and Intrinsic Models are compared at different levels. This slope rises from the primary level toward the higher ones. A slope in the opposite direction is associated with the Ear-born Model while no definite pattern is related to Entity Models.

Figure 6 shows results of students in samples that took both air and wall context if all students are taken together and models are grouped in similarity clusters. When results from the two contexts are combined, it is done in a way that weighted averages are found for each of the models in each of the contexts within the sample.

Upward slope is here again clear for the correct side of models as well as the downward slope for the Ear-born Model. The percentages of students who use Entity Models at these three levels are strikingly similar (52.04%, 52.44% and 52.71%).

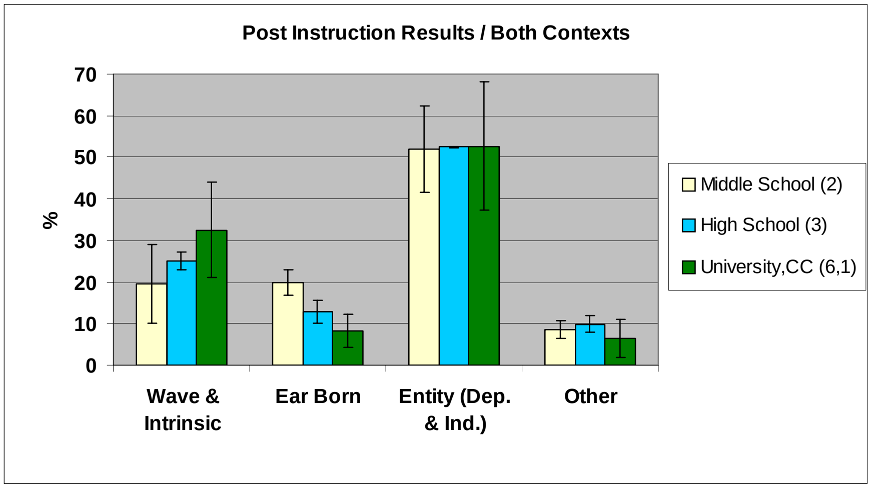

Figure 6. Comparison of post-instruction results at primary, secondary and tertiary levels as obtained by both (air and wall) contexts of the survey (grouped models).

These results show that the test reliably measures students' progress in terms of their usage of correct models (and models that are close to correct). Ear-born motion of sound is less popular at higher than lower levels and the Generic Entity Model (Dependent and Independent) is very stable and on average does not change much with educational level.

Stability of results within the same institution

Another way to determine if results are distributed in a meaningful way is to look at the difference between results obtained from students at the same institution who are enrolled in the courses at different levels. For this purpose, students enrolled in concept-based, algebra-based and calculus-based introductory physics courses at Kansas State University were sampled. The expected result was that the students enrolled in the calculus course will have the best results and students enrolled in the concept-based course the worst results. The obtained results were in accordance with these expectations as can be seen when results of these groups are compared. In the case of the correct model there is a rising pattern that starts with the lowest level course and in the case of the most incorrect model (Independent entity) there is an opposite trend. Models in the middle of the scale are not consistently different.

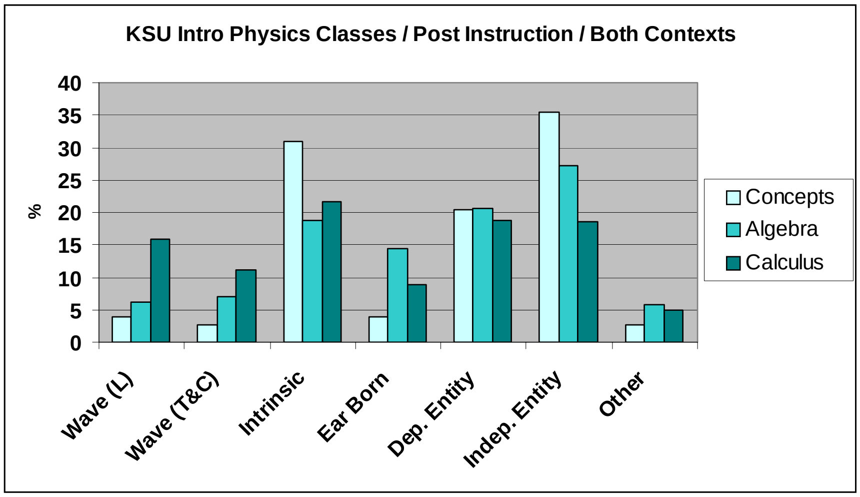

Figure 7. Comparison of post-instruction results at Kansas State University in spring 2003 as obtained by both contexts.

Table 2. shows results related to model distribution numerically as well as results that pertain to self-consistency of students in these different classes. As shown in the table, differences in self consistency (with respect to Wave or all models) follow the same pattern as distribution of Wave Models.

Table 2.

Comparison of post-instruction results as obtained by both contexts at the same institution (KSU) in Spring 2003 and from classes at different levels (in percentages).

ourse Math Level	**N**	**Consistent (Pure model state)**	**Consistent Wave**	**Wave (L)**	**Wave (T&C)**	**Intrinsic**	**Ear-born**	**Dependent Entity**	**Independent Entity**	**Other**
Calculus	126	24.60	15.08	15.87	11.11	21.63	8.93	18.85	18.65	4.96
Algebra	207	15.46	5.80	6.28	7.00	18.72	14.37	20.65	27.29	5.68
Concepts	38	0.16	0.03	3.95	2.63	30.92	3.95	20.39	35.53	2.63

Stability of results across different institutions at the same level

Table 1 and Figures 4 and 8 show that for all models except the correct one, standard deviations between the samples are relatively small when compared to averages. This shows that samples that were analyzed are not very different from each other. This is especially true when one takes into account the relatively small number of the samples and that some of the samples had less than 30 students.

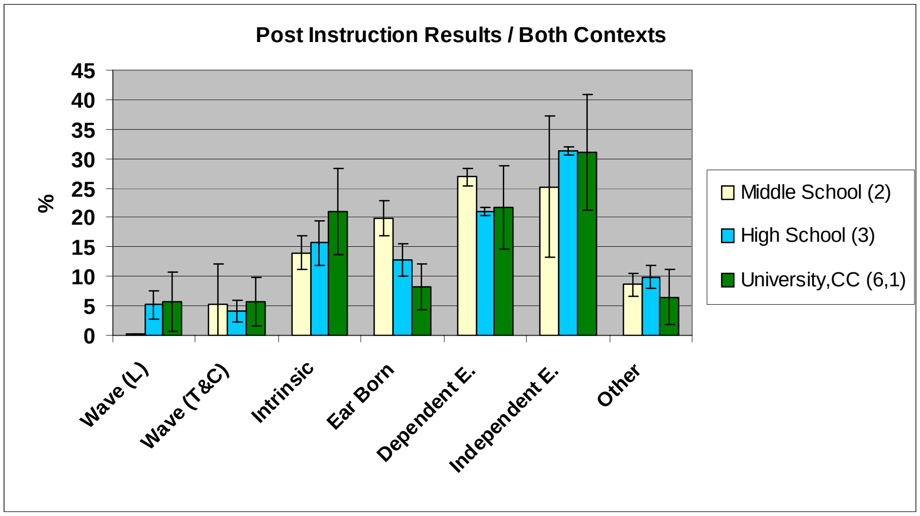

Figure 8. Comparison of post-instruction results at primary, secondary and tertiary levels as obtained by both contexts of the survey together.

The most striking resemblance of the results from different institutions at the same level was obtained in the case of the high schools. In the spring semester of 2003 the test was administered after instruction to students at high schools in Kansas, Minnesota and Croatia. None of the teachers knew about the test during the lessons on sound. Both contexts were administered in each of these samples. In the case of the Croatian sample, all students took both contexts of the test and in the other two samples each student took one context. The middle column of the three columns that represent different school levels in Figure 8 shows these data graphically. Standard deviations between Dependent and Independent Entity Models as obtained from these three schools are the smallest of all models (0.67% each) and the greatest Standard deviation is related to the Intrinsic Model (3.73%).

Small and relatively small standard deviations between samples at the same levels imply that on average, distribution of students' models is rather predictable, i.e. based on these averages and standard deviations, a teacher at any of the levels can pretty accurately determine what he or she can expect in his or her classroom. On the other side, it is possible that the testing itself, in a proposed formative way, may have important instructional value that is worth the time investment in the classroom.

Difference between pre– and post-instruction test results

Another meaningful pattern of differences obtained in the survey is that in all of the cases when the test was administered both before and after the instruction, post-instruction results were better than pre-instruction results. This pattern shows that the test is sensitive to the instructional changes. For the purpose of accurate measurement of the pre- and post-instruction differences, each of the samples that were tested in these two instances is separately analyzed and the results are shown in Table 3. The difference is presented in terms of the gain (percentage increment of the correct answers) and the normalized gain. Normalized gain is the percentage gain achieved divided by the total possible percentage gain or: Normalized Gain = (post-test% - pre-test%) / (100% - pre-test%)

Hake (Hake, 1997) argues that a normalized gain is an accurate measure of the effectiveness (or non-effectiveness) of a particular presentation style. Hake's average normalized gain is usually referred to as the Hake Factor, h.

Table 3.

Results of pre- and post-testing.

Institution	Course Math Level	Context	N Pre	N Post	Method	Pre Test Result (%)	Post Test Result (%)	Gain (%)	Normalized gain
BOTH CONTEXST (WHOLE CLASS)									
University, NY	Calc.	Air	100	95	Research based	9.50	29.21	19.71	0.2178
Middle S., HR*	Algb.	Air	75	99	Research based	0.00	19.19	19.19	0.1919
University, PA	Algb.	Both	12	10	Lec./Demo/Lab	0.00	12.50	12.50	0.1250
High S. (1), HR	Algb.	Both	49	51	Lecture / Demo	0.51	11.76	11.25	0.1131
University, NC	Calc.	Air	57	19	Research based	0.44	9.21	8.77	0.0881
University, KS	Algb.	Both	175	177	Lec./Demo/Lab	2.57	5.79	3.22	0.0330

*Data collected during post-survey phase of the research (all other displayed data was collected during the survey phase)

The model distribution of the sample that had the highest gain looked (before and after instruction) as shown in Figure 9. Air context was administered to this sample both before and after instruction.

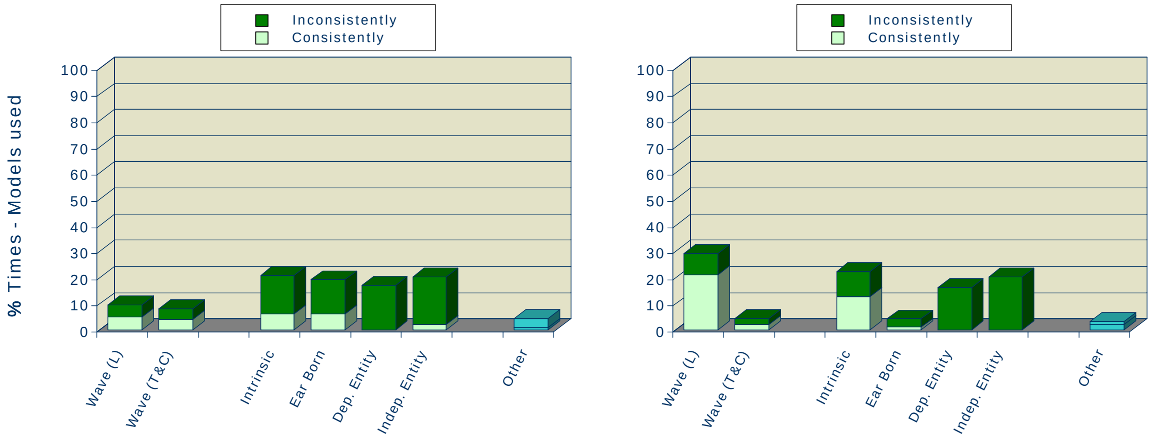

Figure 9. The model distribution of the sample that had the highest gain before (left figure N=100) and after (right figure N=95) the instruction.

The corresponding graphs that show the dynamics of the particles of the medium are shown in Figure 10.

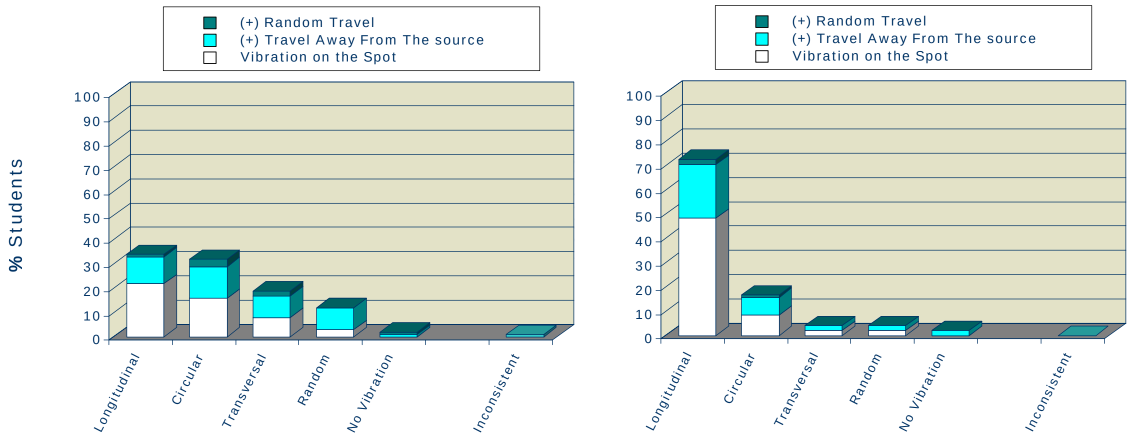

Figure 10. The movements of the particles of the medium expressed before (left figure N=100) and after (right figure N=95) the instruction in the sample that had the highest gain.

Finally, Figure 11 displays change in students' model states in this and sample:

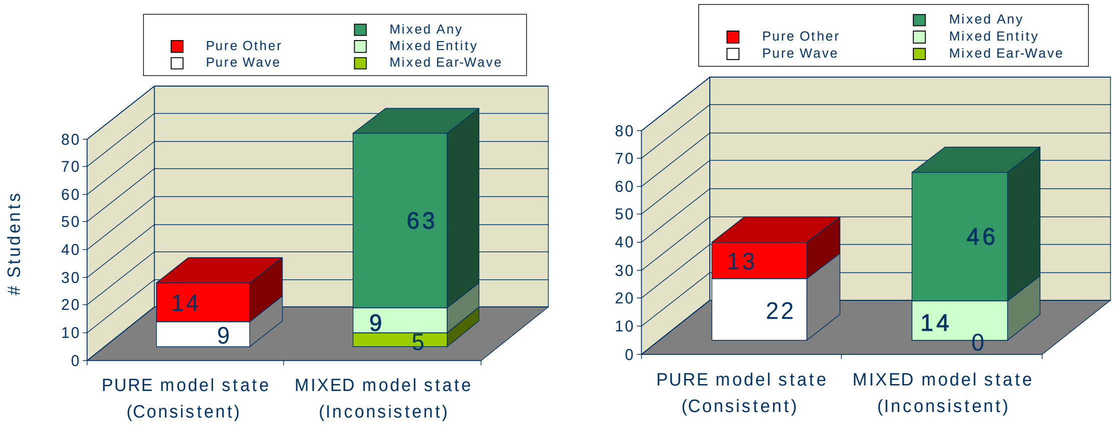

Figure 11. Students' model states before (left figure N=100) and after (right figure N=95) the instruction in the sample that had the highest gain.

Appendix B: Test Validity and Reliability

While discussing test validity and reliability we will combine results obtained during the survey phase of research with those obtained during the post-survey phase. The reason is that the survey phase was by large the most extensive part of the research. However based on the results in survey phase of research two of the answer choices were improved. The test was than revalidated in the post-survey phase and the positive changes where demonstrated but not at such a large scale as during the survey phase. Therefore we will demonstrate that the test was reliable and valid instrument during the survey phase and then even further improved during the post-survey phase.

Two primary ways in which we validated the test during the survey phase of research was through interviews of 17 students and through correlation analysis of answer choices obtained from a sample of 1600 students at 14 tertiary institutions, 4 High Schools and 2 Middle Schools.

Validation through the interviews

Seventeen KSU students enrolled in all of the three levels of introductory physics courses participated in validation interviews. Students' models were determined through the open-ended questions and compared to their answer choices in the test. Open ended questions were administered either before, after or during the test (as part of the think-aloud protocol). In many cases (13/17), as the last thing in the protocol, students' models were additionally discussed based on previously prepared graphical representations of the models for additional verification.

In this protocol, we considered the probe of a model invalid if a student, for whatever reason, picked the choice that did not correspond to the model that he or she was expressing verbally. The invalid probe of the model could have happened four times in each of the tests because there are four model-defining triplets in each of the six-question tests. With 17 interviewed students this makes total of 68 model probes. Of these, six probes (i.e. 8.8%) were deemed invalid based on the procedure described above. Six invalid probes that occurred were made by six different students. There was only answer choice related to which a pattern of misinterpretation was observed (5a). Three students misunderstood this choice, all in the same way. One student misinterpreted choice 6a and remaining two invalid probes occurred not because an item was misinterpreted, but because the statement was misread. These two students noticed their “mistake” in second reading and corrected themselves. These results on their own indicate that even with the two somewhat problematic answer choices the survey version of the test can be considered valid in more than 90% of instances.

Correlation analysis of answer choices

As a quantitative complement of the validity verification through the interviews, correlation coefficients between all of the answer choices were calculated using data on taint from the previously mentioned, large sample. Number of analyzed surveys is greater than number of students because some students tool both, pre and post test.

We wanted to determine whether students who choose multiple models in the test do so because they are not sure about the model (no model state), because they like more than one model (mixed model state) or if this happens because of the validity issues with the answer choices. A particular student who is not firm about his/her model or uses multiple models simultaneously may select choices that correspond to different models in different questions. However, the rationale for correlation analysis was that even if many students are not in a pure model state, if a large sample is taken, the answers that are related to the same model should not have negative correlations. Another indicator of possible problems in interpretation would be a significant correlation between the answers that correspond to different models. Finally we expected that stronger correlations should be found between answer choices pertaining to the correct model because students who know the correct model should be less insecure (about their model) than those students that have no formal knowledge on the topic (about their models).

These main points (mentioned above) that we were primarily interested in with respect to the correlation factors are summarized in Table 2. Table 2 also shows correlation factors that pertain to the correct model. High correlations indicate that those students who have the correct model "know what they do" and are not "mixed" a lot. Indicators in Table 2 add to the quantifiable results that help to determine possible issues in test validity, but they are also very useful in determining the applicability of the test at a specific level. All of the data presented in Table 1 were collected in 2003. For the purpose of determining these correlations, results from the pre- and post- instruction tests were taken together but sorted out with respect to the context. The indicated level of significance was chosen as 5% with respect to significant positive correlations between different models to grasp broader into possibly problematic items.

Table 2

Identifying possibly problematic answer choices through correlation analysis of the choices – survey results

School level		**Tertiary**		**Secondary**		**Primary**	
Context		**Air**	**Wall**	**Air**	**Wall**	**Air**	**Wall**
N		**1132**	**429**	**185**	**115**	**64**	**68**
Desirable correlations related to the correct model between relationship defining questions (Q1, Q4, Q5, Q6) (6 possible)	Correlation of correct choices is highest in respective question	**6**	**6**	**6**	**6**	**1**	**6**
	Sig. at 5%*	6	6	6	6	0	5
	Sig. at 1%*	6	6	6	6	0	4
Desirable correlations related to the correct model between all questions (15 possible)	Correlation of correct choices is highest in respective question	**15**	**15**	**14**	**15**	**2**	**9**
	Sig. at 5%*	15	14	13	13	0	6
	Sig. at 1%*	15	12	12	11	0	5
Problematic correlations between relationship defining questions (Q1, Q4, Q5, Q6) (180 possible)	Primary choices related to the same model with negative correlations**	1 (1c-5a)	1 (1c-5a)	1 (1c-5a)	1 (1c-6a)	13	8
	Significant positive correlations between different models (at 5% sig. *)	8 Dep. & Indep. models	0	3 Dep. & Indep. models	0	7 Various models	6 Various models

* Two tailed

** In counting these instances we ignored situations when the primary choice (or their sum) was negative but the secondary choice in the question was the one with the highest correlation among those in the particular question.

Table 2 shows that at the university level there is one instance in each of the test versions (air and wall contexts) in which two primary choices are negatively correlated. In both cases this is between choices 1c and 5a. That negative correlation shows that students who have the model associated with choice c in question 1 (Dependent Entity Model) will, in principle, avoid what we considered the corresponding choice in question 5. This result perfectly corresponds to our findings in the interviews. In the interviews we identified the nature of the problem and through correlation analysis we identified that the problem is present at a large scale. Due to the insight obtained through the interviews and related to the nature of the problem, we were able to address the problem in the final (post-survey) version of the test.

With respect to the strength of the correlations pertaining to the correct choices, university students had a perfect score in both of the contexts. All combinations of choices pertaining to the correct model had the highest correlations among the choices in respective questions. Also, all correlations between the relationship defining questions (Q1, Q4, Q5, Q6) were significant at 1% level two tailed. If dynamics defining questions are considered as well (Q2 and Q3), all correlation coefficients between the correct choices were highest of all in a particular question and most of them, although not all, were also highly significant.

An unexpected result that Table 1 shows with respect to the university students is related to the number of significant positive correlations that pertain to different models. However, in each of these cases mixing occurred only among Independent and Dependent Model choices. However these two models do not have a firm boundary and from this perspective this result is not worrisome. In addition, we observed that that these two models may hybridize during the test taking into a model that is a combination of the two (sound starts as an independent Entity, shakes the medium and than shaken medium transfers the sound further). When this happens a student may pick choices pertaining to either of them. Finally, unlike in the case of the correct model which is used by students who likely know what they are doing, the Dependent and Independent Models are at the bottom of the correctness scale. Students at this end can not be expected to have as stable ideas as those who have the correct model. Combining these arguments with the results from the interviews gives a solid ground for the claim that Dependent-Independent mixtures indicated in the right-most column of Table 1 reflect valid mixed states. The program for model analysis of the test results sorts out students that use a mixture of dependent and independent entities exclusively. Surprisingly, these mixtures are not pronounced in the wall context of the test at all.

Analysis of the same data with respect to high school students reveals similar issues. All that was said related to correlations between the choices at the university level applies here too. When middle school students are considered, the results show evident need for further study on applicability of the test at this level. Some encouraging results were obtained in the spring semester of 2004 when Middle schools students showed impressive gain (larger than most of the university samples).

Post survey test modifications and validations

In the post survey phase, issues that were identified in the survey testing were addressed and validity of the new version was verified again. The test choices were improved based on the qualitative and quantitative results that were collected in the survey phase and based on the direct suggestions that students gave during interviews. The modifications were made primarily to address the problem with answer choice 5a but question 6 was also modified to avoid possible issues with understanding of answer choice 6a.
This new test version was additionally validated in three ways: (1) Through the verification of the positive change of earlier problematic correlations (N=339), through expert reviews and through role-playing validation in which (another set of) experts in physics assumed the roles of students having models that the test probes for and who took tests that way.

When the test was once again administered to students, obtained results showed favorable change in the correlations of answers 1c-5a and 1c-6a that we were aiming toward. Also, there was only one significant positive correlation between different models (and this was combination of Dependent and Independent Models, which we showed, is not a validity issue). The result of the role-playing validation was that all of the experts straightforwardly picked choices that were corresponding to the models they were "assigned to".

These results show that in the post survey versions of the test, weak points (choices 5a and 6a) of the survey test version were addressed, while other relevant parameters stayed the same as in the survey version and no new problematical issues arose. This gives ground to use the results obtained with the survey version of the test (8.9) as a basis for conclusions about the validity and reliability of the final version of the test (9.2). It further makes plausible the claim that, had the final version of the test been administered to a large sample as the survey version was, the results would have been the same as or superior to those of the survey version of the test.

In addition to meaningful correlations between the answer choices (at the secondary and tertiary levels but not at the primary level) reliability of the test was shown through:

1. Stability of the results across the different institutions at the same level as reflected through the small standard deviations around the average percentages at each of the models is represented in each of the samples.
2. The expected direction of differences between results in terms of the usage of the correct models and in terms of the students' self-consistency. Correct models and self-consistency are more frequent among students:
   - at higher educational levels than lower,
   - in more advanced introductory physics courses at the same institution than on the lower ones, and
   - after the instruction than before it.

There are four threats to reliability of the test and results shown above demonstrate that this instrument is resistant to each of them. Meaningful correlations between the answer choices indicate that content sampling error is not an issue in this test. The content sampling error is further reduced by probing a single model multiple times in this test. The second and third reliability indicator listed above show that the test is resistive toward the occasional sampling error. Examiner error, the third of the four reliability threats, is not measurable and it was reduced through the standard introduction. Finally, the scorer error was not an issue at all because of the computerized analysis of the results. This closes the list of threats to the test reliability. Because all four of the threats to the reliability of the test were well addressed in the study, this gives ground for the claim that the test is a reliable instrument.

In addition to earlier described interview protocol and correlation analysis, we employed several other validation procedures. Of these we will mention experts' review of the continent and correctness of the answer choices and instructional sensitivity of the test.

Experts' review of the content and correctness of the answer choices

A panel of experts (Ph.Ds in Physics) reviewed the test in two phases of its development. The first time was at the end of the pre-survey phase before we administered the test to a large sample. The second time was in the post-survey phase after we made modifications based on the results in the survey phase. Each time four experts reviewed the test to determine if choices that we consider correct are (1) correct and (2) the only correct answers and to give us feedback on the clarity of the sentence formulations in the test. Their suggestions significantly contributed to quality of the test and their verifications of correctness of a single choice in each of the questions strengthened the case for validity of the instrument.

Validity of the test at Middle School level

Table 2 in the Appendix A shows a number of undesirable correlations related to the survey data collected at the primary level (which is not the case with the secondary and tertiary level), especially in air context. This may be an indicator that the test might be demanding for this age and therefore not applicable. The situation was not much better when the test was administered to middle school level students as a pre-instruction test during the post-survey phase. However, when middle school students took the test after instruction (which aimed at eliciting alternative mental models), the improvement with respect to usage of the correct model and students' self-consistency was surprisingly large. The percentage of students who consistently used a model increased from 1.33% (before the instruction) to 9.9% after the instruction. More importantly, the Longitudinal Wave Model became the most frequently used model of all, with 5.5% students using it consistently (out of 19.19% total). Figure 12. shows the model change as obtained from this sample before (N=75) and after the instruction (N=99).

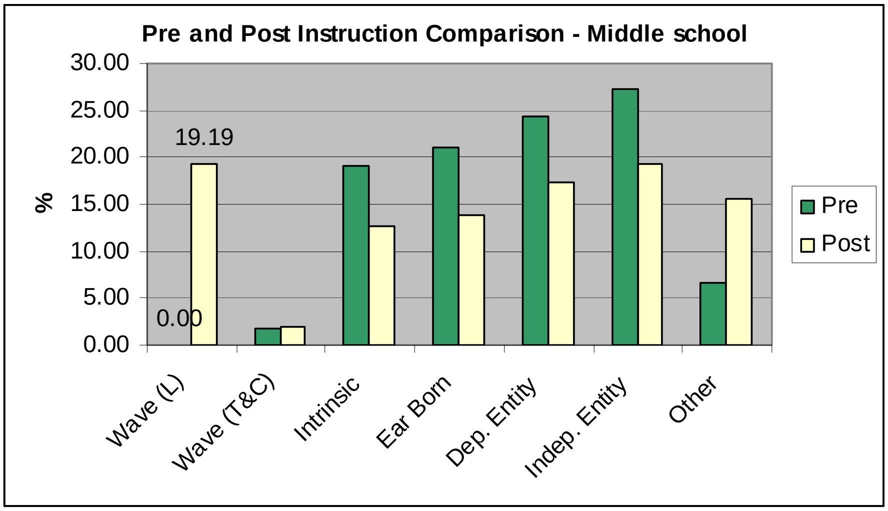

Figure 12. Model change at the middle school level as obtained after model-targeted instruction.

The learning gain obtained from this middle school sample (19.19% unmatched and 18.57% matched) was one of the highest observed in this study. These results show that the test might be applicable also at the middle school level in some form, but more research is needed related to this. Another reason for not abandoning the middle school level too soon is the fact that correlations at this level were based on a far (roughly 10 times) smaller sample than for the college students.