2006 National AAPT Meeting – Summer
Notes from July 24-27
Syracuse, NY

The following notes were taken by Dr. Van Domelen during the regular sessions of the 133rd National Meeting of the American Association of Physics Teachers.

<u>Contents</u> <u>Page</u>

Contents	Page
Table of Contents	1
Glossary of Abbreviations	2
Airing of Grievances	3
Monday, July 24	
AK: Assessment Instruments and Teaching/Learning Theories	4
BB: Plenary II: Pulsars and Extreme Physics	8
BH: PER from Different Perspectives: Why Bother? (Invited Poster)	10
Tuesday, July 25	
CM: PER in Introductory Physics Courses	13
Labs Committee Meeting	18
(Did not attend evening session)	
Wednesday, July 26	
EE: Quantitative Assessment of Cognitive Modesl (Invited)	20
FG: PERC Bridging Session (Invited)	23
FI: Bridging Multiple Representations: From Cognitive Science to Physics Education (Invited, partial)	25
Thursday, July 27	
TP-A: Issues in Concept Inventory Development	28
IT: PERC Invited Talks	30
RT-A: Roundtable, Cross-pollination in Science Education Research	34

Abbreviations

There are a few concepts that came up frequently enough in the talks I attended that I gave them their own abbreviations. (Not all will appear this time; I just pasted in the 2004 glossary and added to it as necessary.)

ADT – Astronomy Diagnostic Test

BCI – Biology Concept Inventory

CSEM – Conceptual Survey of Electricity and Magnetism

ECR – Elicit, Confront, Resolve. The UWash method.

FCI – Force Concepts Inventory, a common mechanics test.

FMCE – Force and Motion Concept Exam, a common mechanics test. I may sometimes type it as FCME.

GG – Gender Gap, a disparity between male and female results, usually to the detriment of the women.

IE – Interactive Engagement, one of many strategies that get students more involved in their own learning.

ISLE – Investigative Science Learning Environment

MBL – Microcomputer Based Labs

MPEX – Maryland Physics Expectations Survey, a test of student expectations about physics and physics courses. MPEX2 is a recent revision.

N1L, N2L, N3L – Newton's First Law, Second, Third.

PRS – Personal Response System, a means of letting students answer multiple choice questions in class using a remote control. AKA "clickers".

SPS – Society of Physics Students

TIMSS – Originally "Third International Math and Science Survey", but they changed what the acronym means in order to keep using all the logo stuff as they move on. ☺

I will sometimes add personal comments in [brackets]. I will try to withhold value judgments most of the time...if you feel the need to argue a point about a paper, please contact the author, not me (email addresses provided). However, if you are the author and feel I have missed the point of your paper, let me know and I will attempt to remedy the situation.

Airing of Grievances

There were a lot of problems with this meeting, both with the facilities and the actual organization of the sessions. Rather than go into details, however, I will focus on two major problems. I am including these here in my summary because hopefully people who are involved in future meetings will read this and address the issues. Or, at least, hear about this from people who do read it.

1. Lack of advance information about the facilities. Most of the worst issues with the facilities could have been blunted or eliminated if we'd just known more about what was going on far enough in advance to make alternate plans. Complaints about the meal plan, for instance, could have been addressed (or we could have just avoided the meal plan and eaten off-campus) if we'd known before signing up how many problems there would be. Knowing that the dorms were atop a rather large hill and that the shuttle service would only run once an hour would probably have led a number of people to opt for hotels instead. That sort of thing.

Let us know how things will work before the registration deadline, so we can make an actual informed choice, rather than being ambushed by problems after we get to the conference.

2. Poor session scheduling. Having the contributed talk sessions at 8 AM and 8 PM with none in-between was bad for a host of reasons, and sometimes it turned out there were classes scheduled during times that we were supposed to have talks. Poster sessions were noisy and crowded (sometimes violating fire code, as far as I could tell). Everything was spread out all over campus, an issue for anyone who wanted to attend parts of multiple parallel sessions. I suspect part of the problem came from scheduling the conference while classes were in session, leaving us the odd times and bad rooms in many cases (and also leaving us stuck with the worst dorms on campus, as all the closer ones were in use by students).

In any case, I found myself able to go to fewer talk sessions than last year, and the night sessions meant many people had to choose between missing half the talks for the day or getting enough rest to make it to the morning talks. Granted, it was easier to attend committee meetings and crackerbarrel sessions, but that may not be an even trade for most.

Monday, July 24, 2006

AK – Assessment Instruments and Teaching/Learning Theories

[Note: session started late due to problems with connecting laptops to the projector, and several talks were impacted by this problem.]

AK01: Classification of Physics/Astronomy Concept Inventories – Elizabeth Peak, Southern Illinois University at Edwardsville, epeak@siue.edu

Instruments were split into three categories: local tests, efficacy tests and diagnostic instruments (per Lindell and Foster). Local tests have no particular reliability or validity testing done on them, and are created by instructors for personal use. Efficacy tests are researcher-made and probe various concepts, but do not make significant use of interviews to determine the content, so cannot say what students do know, just whether they know the correct things. They are, however, tested for reliability and validity. Diagnostic instruments have all the characteristics of efficacy tests, plus use interview data extensively to determine distracters. With effective distracters, it is possible to diagnose incorrect views held by students, not just ascertain that the students lack the correct view.

Most "concept inventory" tests are really efficacy tests under this taxonomy, and cannot diagnose incorrect student views. Very few tests labeled as diagnostics are actually diagnostic in nature.

AK02: Validity and Reliability of the FMCE – Susan Ramlo, University of Akron, sramlo@uakron.edu

Opened with background on Thornton's development of the FMCE, defined validity types (face, content, criterion, construct) and then said content validity would be focused on here. Also defined reliability using the alpha measure (the probability that a false effect would be rejected by the instrument).

Initial FMCE work with small N had good confirmatory factor analysis showing the clusters of items seemed to actually cluster the way they were intended, and had an alpha of 0.91, demonstrating good reliability. However, while alpha was good, the small N meant that beta (the probability that a real effect would be spotted by the instrument) was too low.

A new study was performed with N=146, to bring beta up to a reasonable level (exactly what beta was did not get discussed, but N=146 gives you a good beta for some fairly small effect sizes). Alpha remained at about 0.91 on the posttest, and there were strong confirmatory factors (cluster of N1L and N2L, another of N3L, another for velocity) on the posttest. The pretest had weaker factors, but students are expected to have answers closer to random on a pretest.

Conclusion: the FMCE has good validity and reliability.

AK03: A Switch Effect in Concept Test Questions – Lei Bao, The Ohio State University, lbao@mps.ohio-state.edu

[Computer issues and a dead laptop battery shortened the talk somewhat.]

"Switch Effect" is defined as a change in student response patterns that takes place when the order of questions is changed. The work in this talk was inspired by Kara Grey's previous work on FCI question order, and used the two "car pushing truck" items on the FCI for its initial study.

Ordering can trigger different ways of thinking, flipping a switch in the subject's mind.

New items were designed to better test this effect after some interesting results on the FCI-item version. Each item had a slightly different problem, and while some shared the same context, others did not. When the two problems in a trial had different contexts, they found no order effect, no "switch". However, when the contexts were the same, the effect remained. It is thought that the relative measure of the switch within a particular context can help measure more subtle effects of student understanding.

AK04: Conceptual Priming in Multiple Choice Questions – Jing Wang, The Ohio State University, wang.870@osu.edu

[This talk went by very quickly, apologies to the author if I wholly misinterpret it.]

"Stem completion" priming presents subjects with a target, then has them retrieve it from partial information. For instance, in a simplistic case, if you say "Stem completion" to someone several times, then later say, "Stem...?" and they respond with "completion," then stem completion priming has occurred.

Conceptual priming, however, uses partial meaning rather than partial form to cue the completion. A cueing page would provide questions aimed at triggering particular ideas (i.e. having students draw a force diagram), then a testing page would ask a conceptually related question (i.e. identify force pairs) to see if the students got it. Testing pages can use either familiar or unfamiliar settings, but should all have a conceptual link to the cueing page.

Cueing and then testing was shown to help in novel settings, but not so much in familiar settings. **Perhaps conceptual priming aids in transfer?**

AK05: Challenges in Developing Effective and Useful TIPERs – Curtis Hieggelke, Joliet Junior College, curth@jjc.edu (With David Maloney)

Opened with a very fast overview of TIPERs (Tasks Inspired by Physics Education Research), then moved on to the specific challenges presented by making them for E&M topics.

Issues in magnetism include: there is little available research, it has a strong component of rules-memorization, it MUST be done in 3-D, there is a weak connection between formalisms and reality, and it's tied to electrostatics (which is very vocabulary- and abstraction-heavy).

10-15 tasks per specific issue were developed for magnetism and 5-10 per issue for electrostatics, and tested them all out. An alternative to the CSM was also developed to better cover the topics the TIPERs taught. mTIPERs and eTIPERs are magnetic and electrostatic TIPERs, respectively.

AK06: The CSEM – Thom O'Kuma, Lee College, tokuma@lee.edu

This talk was essentially a historical overview of the development of the CSEM.

In 1995, a need was seen for an E&M instrument of some sort. There were plenty of mechanics instruments, but almost nothing for second semester aside from some circuit tests.

The E&M Concept Inventory was a first step, but never really intended to be developed into a final instrument on its own. The ECI (Electric Concept Inventory) and MCI (Magnetic Concept Inventory) were developed from this work, later being renamed CSE and CSM in 1997. Items from both surveys were then combined to make the CSEM, with development of that instrument finalized in 1999. The CSEM showed students have trouble with electrostatics, and a LOT of trouble with magnetism.

Now it's time to develop a truly diagnostic instrument.

AK07: Towards a Model of Inconsistencies in Student Responses – Steve Stonebraker, The Ohio State University, sstoneb@mps.ohio-state.edu

Opened with an intro to Bao's Model Analysis wherein students are treated as being in quantum-like states, either pure or mixed.

Students have inconsistencies in performance comparing show-work problems and multiple choice (MC) items. Is there a pattern?

Midterm exam problems of both kinds (MC items being FCI-like) were examined. Students were assigned to one of three states: PI (pure incorrect), M (mixed) or PC (pure correct).

Those who were PI on MC tended to be M on show-work ($N=5$, $p << 0.01$). Those who were M on MC tended to be PC on show-work (stats not as good). There isn't enough data yet to make any strong statements, but it's starting to look like there is a link between the two problem types.

AK08: A Post-Positivist Perspective on Physics Education – Melissa Dancy, University of North Carolina at Charlotte, mhdancy@email.uncc.edu

This talk is in part a teaser for a PERC poster.

Positivism is the stance that there exists an objective reality that we can know, and tends to be the (often unexamined) position of scientists.

Post-positivism posits that even if there is an objective reality, we can't really know it, as cultural bias always kicks in. What we choose to research and how we interpret the results are culturally determined, "ways of knowing" make a difference. Truth is socially constructed, even if it happens to use pieces taken from an objective reality.

Leaving aside the more extreme types of post-positivism where objective reality is not assumed to exist, PER is still much more consistent in practice with post-positivism than with positivism. PER practitioners are straddling a line between the positivist assumptions of physics, and the post-positivist evidence of education research, so we need to find a way to work in a mixed mode, assuming there's anywhere that the two views can overlap. [Editorial comment: dualistic thinking is unfortunately all too common, and middle grounds are often attacked from both sides on the "if you're not with us, you're against us" principle. Trying to find a way to accept objective reality

AND socially-constructed knowledge may just get us dismissed as unscientific by the physicists and dismissed as dogmatic by the post-positivists.]

AK09: Representation of Diagrams in Memory After Short Exposures – Adam Feil, University of Illinois at Urbana-Champaign, adamfeil@uiuc.edu

The main question being asked here is, "How can we know what is 'seen'?" We know experts and novices classify things differently (novices by surface features, experts by underlying structure, in general), but there are confounding effects in measuring this. One such effect may be disentangled by looking at diagrams.

We better remember the parts of a picture that we pay attention to, and the sorts of things a novice attends to should be different from the things an expert attends to.

Pairs of images were constructed that differed in some aspect, either a physically important difference or one that shouldn't affect the physics.

Example 1: A modified Atwood where blocks on opposite sides of a triangular peak are connected by a rope over the apex. In one case, they're at the same height, in the other case they are not. This is a difference that doesn't affect the physics.

Example 2: A battery, bulb and two wires. In one case, the one wire goes to the base of the bulb and the other to the side, so it can light up. In the other case, both wires go to the base, so the bulb will not late. This is a difference that affects the physics.

Subjects are shown one of the pictures in a pair, and then later either shown the same picture or the other picture from the pair, and asked if they've seen it before. Then they are shown both side-by-side and asked if the two are different in an **important** way (without asking if either is "right or wrong"). They are asked to briefly describe the differences. This all takes place on a computer screen, so they type in their responses.

Experts should spot non-trivial differences more frequently than trivial ones, while novices should be equally sensitive to all differences. Data collection had only started a week before the conference, so there are no conclusive results so far, but novices do seem to be focusing on surface features in interviews (i.e. the lower block in a modified Atwood must be heavier, and therefore sliding down).

Not knowing what features are important and which ones can be ignored is likely to increase cognitive load.

AK10: Student Resources for Understanding Motion – Brian Frank, Arizona State University, bwfrank@asu.edu

Resources are chunks of knowledge that students use for reasoning. As used in this talk, they are essentially "P-prims", and the examples given were:

Speed/Time: faster means less time

Speed/Distance: faster means farther

Distance/Time: farther means more time

Students are given different prompts in an attempt to activate different resources. Commonly a single prompt is all they need, but sometimes the subjects lack a particular resource and require multiple prompts to build it from the ones they do have. For instance, using Speed/Time and Speed/Distance to work out Distance/Time. These serial applications are not always done correctly, of course.

Parallel applications are also sometimes used, such as a case where something goes farther and is faster, so students presume the time taken is unchanged.

Resources may be correctly or incorrectly applied, but they are not inherently right or wrong.

AK11: Manipulating Mental Images and the Application of the Biot-Savart law – Ramon Lopez (with Kastro Hamed), Florida Institute of Technology, relopez@fit.edu

When using a 2-dimensional representation (i.e. a drawing) of a 3-dimensional concept (such as magnetism), problems ensue.

Magnetosphere Substorms (Auroras) were used as an example of a 3-D magnetic/current system, where current of ions in the tail of the magnetosphere divert inward to the ionosphere and form aurorae.

2-D representations were contrasted with stereograms (paired images on a computer screen viewed through special frames to present a different image to each eye and therefore create the illusion of depth). The stereograms were also interactive, and the students could rotate the images on the computer screen.

Upper-level students at the University of Texas at El Paso were interviewed. All of them had demonstrated their ability to use the formal Biot-Savart law in terms of solving the equations. Pretty much everyone given the 2-D representation failed to correctly solve the problem, while everyone given the 3-D representation was successful. Many students misinterpreted the 2-D picture, and the ability to rotate the 3-D image was explicitly mentioned by students as a reason for their success.

Having to not only construct a 3-D image from a flat one, but also having to mentally rotate it, increased the cognitive load on the students. The 3-D simulation provided a sort of external cognition, and the manipulability may turn out to be more important than the depth or lack thereof. Future work will explore this.

BB – Plenary II

BB01: Pulsars and Extreme Physics – Jocelyn Bell Burnell CBE, Oxford University, Jocelyn@astro.ox.ac.uk

Professor Bell Burnell discovered the pulsar, or pulsating radio star.

The first pulsar observation was mistaken for an equipment error and not logged. If the "error" had been logged, the observer would have gotten credit for the first observation, so be sure to log everything!

A pulsar is a neutron star that rotates and emits an energetic jet off-axis that sweeps around with a frequency of around 11 Hz. Not all neutron stars are pulsars, but all pulsars are neutron stars. Pulsars are one of the later stages in stellar life.

Stars begin their lives in nebulae, where perturbations in the gas form the seeds for stellar birth. They also tend to end their lives with nebulae, casting off material either gradually or explosively to form "ring" nebulae (which are actually spherical shells, it's just that they're so thin we can only really see the edge-on parts). Most stages in stellar evolution can be found in or near the constellation Orion.

The remnant of Supernova 1987A has become a neutron star sitting at the center of a ring nebula. Such nebulae cast off by supernovae tend to last on the order of 10^5 years.

Orbital junk is making deep field astronomy increasingly difficult in visual wavelengths.

Some properties of pulsars:

- mass of about 10^{30} kg.
- radius about 10km.
- Density approximately equal to that of a nucleus, hence "neutron" star.
- Period of pulsing is equal to the rotational period, which is usually from 1.6ms to about 8s, although outliers exist.
- One of the fastest pulsars has a period of 1.5578064488725 ms, with an error of 0.00000000003 ms.
- The most perfect circle in the universe is a pulsar orbit. With a radius of about 10^8 meters, it has a variation from circularity of about 40μm.
- There is an atmosphere of accreted materials 1-2cm thick.
- Gravity at the surface is strong enough to bend light about 20-30^o over the horizon and generate a noticeable redshift near the surface. The value of γ is about 2 for purposes of time dilation.
- The tidal gradient is truly nasty, leading to "spaghettification" of anything too near the surface.
- The pressure at the center of a neutron star is about 10^{24} atmospheres.
- Typical magnetic fields are about 10^8 T, or a trillion times that of the Earth. Typical electric fields are about 10^9 V/cm. Combined you, get a Lorentz force that is **very** large.
- A "neutrosynchronous" orbit at a radius of 5000km would involve moving near the speed of light, for typical rotational periods.
- Robert Forward's Dragon's Egg involves life on a neutron star. (Note: if you want to search for works by Robert Forward, look only for text SF. His son Bob writes SF for television.)

The magnetic axis of a pulsar is typically not parallel to the rotational axis, resulting in a cone of magnetic fields that create a sweeping radio beam. We can only detect pulsars **as pulsars** if the beam sweeps across us, otherwise at best we can detect them as neutron stars. About one in five pulsars are estimated to be pointed the right direction. This may seem high, but the radio beam is fairly wide for most pulsars.

The radio beams are relatively weak, so we can only really detect pulsars on our own side of the galaxy. But the high mass makes them insensible to most perturbations, so they are estimated to have a "wobble" of less than one part in 10^{21} in their oscillations. There is no other timekeeping system that even approaches this level of accuracy, it is inferred by how often a timekeeper of known accuracy disagrees with the pulsar.

Many pulsars are in binary pairings with either regular stars or (in rare cases) other pulsars. General Relativity predicts that the oscillations of these masses as they orbit each other should give off gravity waves, with pulsars making excellent sources of gravity waves. Since any sort of wave emission takes energy away, the binary pair should slowly get closer together, moving faster and faster before finally merging.

Since pulsars are such good timekeepers, we can use the rotation of a pulsar to measure the rate at which the revolution of the pulsar changes and see if it is indeed speeding up. 30 years of observations made on the first known pulsar-star binary system agree with GR's predictions to within 0.1%. Recently, a pulsar-pulsar pair has been found, and is expected to give even better data.

Pulsar timekeeping has also been used to find planets around pulsars or around pulsar-star binary systems. The slight Doppler shifts caused by a nearby planet are easily measured against the ultra-stable background frequency.

The Pulsar watch company tried to sue astronomers over the rights to the trademark on "pulsar". They lost. [I guess they tried for a too-wide claim on the name, I'm sure they could easily defend a trademark on using the term as a brand of watch, since while astronomers use pulsars as timekeepers, they don't actually sell pulsars.]

BH – Physics Education Research from Different Perspectives: Why Bother? (Invited Posters)

BH01: Seeing What Works: An Empirical Approach to PER – John Thompson, University of Maine, thompsonj@maine.edu

This poster presented exemplars of various approaches. The goal was to improve curricula and student learning. The presented approaches proved to be very effective in the long run if allowed to reach critical mass.

The approaches were context specific, but were often (not always) generalizable. They were pragmatic, directly applicable and immediately useful.

Some student difficulties illuminated by this work include:

- Confusion between a thing and the rate of change of that thing (velocity and acceleration, flux and change in flux, etc).
- Failure to properly analyze variables in a multivariate problem.
- Tendency to use "compensation" arguments, wherein two changing quantities are assumed to exactly cancel each other out, rather than realizing that one may be dominating the situation.
- Tendency to take complex representations too literally.

[Note: this was actually the last poster I went to, my hand was hurting and the poster was very dense, so I didn't get as much from it as I would have liked.]

BH02: Breaking the Paradigm: Working from a Sociopolitical-Critical Perspective – Melissa Dancy, University of North Carolina at Charlotte, mhdancy@uncc.edu and Noah Finkelstein, University of Colorado.

Education is necessary in order to improve society. The current system is an impediment to improving society. We need to develop ideas for reform and provide a vision.

"Professor Project" studies attempt to change institutions to make it easier for true reform to happen. We need to either create new systems or at least act outside of the existing systems, as the existing systems are an impediment to change. We need to look outside of traditional roles, and act as a partner to the community, empowering them.

We need to ask different questions, see different answers, and use different methods...but this brings about its own limitations, such as figuring out how to evaluate success. (Also, working outside the system can make it hard to get published or tenured, which long-term acts against any attempts to reform...if the reformer is not heard or hired, it's hard for reforms to happen.)

An example of a different question is to ask "**should** we teach gravity?" instead of the current practice of asking things like "how should we teach gravity?"

The educational system is a staunch supporter of the status quo, so is very hard to change. We need to examine hidden agendas and look at paradigms from the outside.

BH03: SWOSing, and Theoretical Perspectives That Can Explain It – Valeria Otero, University of Colorado, valerie.otero@colorado.edu

[Not the title given in the Announcer]

SWOS = See Which One Sticks. This is the practice of shotgunning terms and concepts in an attempt to see if any of them work for the problem at hand. When students are SWOSing (pronounced "swoozing"), they bring up multiple terms in a short period.

A six-step process of model-development for a concept in magnetism was looked at in the context of SWOSing, to see which frameworks best explained what was going on. These frameworks were Framing, Vygotsky's Socio-Cultural, and cognitive conflict (I think...I didn't write down the third one).

Use of only one term or an otherwise very small set of terms indicated a model had been settled on, and the terms linked well to that model. SWOSing mainly happened in transitional periods, where a model had not yet been settled on.

Instances of SWOSing were linked with mainly evidence-based reasoning, while non-SWOSing periods were associated with all sorts of reasoning styles.

Different research perspectives lead to different conclusions, however. Some parts of the situation were better explained by a certain perspective, and some elements were interpreted in contradictory ways by the different frameworks. For instance, from the Framing perspective, SWOSing is bad framing, but from the Vygotskian perspective, SWOSing is a bootstrapping endeavor, and in something like Model Analysis (Bao) it could be evidence of a mixed state. In other words, SWOSing could be an indication that the students are wasting time running down blind alleys, it could be a vital part of constructing understanding, or it could be evidence that students are moving in the right direction from "pure incorrect" to "pure correct". Even within a framework, contradictory interpretations are possible, such as the view that SWOSing may be a way in which an incorrect framing is corrected.

A conceptual framework to unify all of this is definitely needed.

BH04: Transfer of Learning: Implications for Research, Curriculum Development and Instruction – Sanjay Rebello, Kansas State University, srebello@phys.ksu.edu

This poster presented the current form of the model of transfer being developed at Kansas State (email Sanjay for a copy of the current diagram), as well as work being done in adding teaching interviews to the instructional design process.

The standard instructional design process is a feedback loop involving clinical interviews, curriculum design and field use of the curriculum. Teaching interviews are added to this loop between the clinical interviews (which mainly look at student difficulties with the content) and the curriculum design, in order to help work specific methods of teaching the content.

In discussion, a method called RBC was brought up as used in Math Ed, and was described as similar to APOS (Action Process Object Schema) as per Dubinsky.

[I did not stay for the discussion session, wanting to be sure of catching the last shuttle bus up to the dorms.]

Tuesday, July 25, 2006

CM – PER in Introductory Physics Courses

CM01: Why Do Good Students Fail Introductory Physics, and What Has Math Got To Do With It? – Eugene Torigoe, University of Illinois at Urbana-Champaign, torigoe@uiuc.edu

The researcher was interested in seeing how students understand the meaning of symbolic expressions, so looked at two final exams that covered the same concepts (largely kinematics). Final 1 was mostly numeric problems, while final 2 had symbolic problems (i.e. the givens were not provided as numbers, but as things like v_1 and t_2). The averages were 85% and 57% respectively, and the actual spread could be even worse.

Symbolic manipulation is not the same as numeric computation as far as students are concerned.

10 students were interviewed three times over the course of the semester, working through exam-like problems in a talk-aloud protocol. The students were drawn from a wide range of course grade levels. Students who were unable to solve a problem symbolically were generally able to solve it if given numbers to plug in. Some students claimed that with numbers they were "actually solving something" (and could use a calculator).

[Talk ran out of time.]

CM02: Can Students Evaluate Their Own Understanding? – Jessica Watkins, Harvard University (Mazur group), watkinsj@deas.harvard.edu

Students seem, anecdotally, to not realize when they understand or misunderstand something. When they seem confused, this is when they actually understand more. "To wonder is to begin to understand."

Students were asked to rate their comprehension on a 5-point Likert-like scale throughout the course of instruction using methods that included peer instruction and Just in Time Teaching. This self-rating was compared to direct assessment of student understanding, such as exam scores.

The correlation between self-assessment and external assessment was only 0.15 initially, but had risen to 0.46 on the third exam. Confidence itself also rose slightly over time. Both effects plateaued, however, and the higher load on the final exam dropped the correlation to 0.29.

The conclusion drawn from this work is that students can learn to evaluate their own understanding, but are not naturally good at it.

CM03: Using Physics Jeopardy Problems to Assess Student Learning – Lili Cui, Kansas State University, lili@phys.ksu.edu [lili@umbc.edu is new address, she is now at University of Maryland, Baltimore County]

This work uses "neo transfer" concepts, in which transfer is more common than in traditional definitions. In particular, uses the Kansas State model of transfer.

Physics Jeopardy Problems (PJPs) are ones in which an expression or equation is presented, and students need to figure out a reasonable situation that would lead to that expression or equation. For instance, "$mg\cos\theta - \mu mg\sin\theta = ma$" might suggest an object sliding down an inclined plane. PJPs prevent the usual plug-and-chug methods from working, and help reveal how students understand expressions. They focus more on process than results.

Semi-structured interviews were performed with students from second semester calculus-based introductory physics to determine if the use of PJPs helped with transfer of calculus knowledge to physics.

If given numerical expressions, students tended to extract variables and constants based on the units, and then use these to figure out the problem type. (i.e. "This problem has kq/r^2, so that means it's an electric field problem.") There was a lot of pattern matching without physical reasoning, often incorrect pattern-matching. (To continue the example, the expression was actually the integral of $kqdr/r^2$, so it was electric potential.)

Students had difficulty understanding the meaning of variables of integrations or differentiation. (i.e. "What does d**s** stand for here?")

Transfer was found to be difficult in this situation.

CM04: From TIPERs to TICERs – David Maloney, Indiana University Purdue, Maloney@ipfw.edu

Looking to expand the TIPERs to Chemistry Education Research. TIPERs have been extensively researched and field-tested, there's no reason they couldn't be adapted to other scientific fields.

The particulate nature of phenomena is the primary conceptual issue for chemistry, along with the macro vs. micro comparison.

TIPERs often elicit "natural" responses rather than book answers, but would that work with TICERs, given the lack of "natural" responses to many of the topics? The case of boiling point alteration with the addition of solutes was considered, using prototype TICERs after traditional instruction. The results were illuminating and useful.

CM05: Instruction Evaluation of Electricity and Magnetism in a Large University – Genaro Zavala, Tecnologico de Monterrey (Mexico), genaro.zavala@itesm.mx

A modified CSEM was used (adding circuit problems), and judged to be a good multiple choice test. Concentration Analysis was used to determine the student state, where the concentration factor indicates what fraction of a group answered the same. Γ in particular is the concentration of wrong answers, and you want it to be low post-instruction.

Gains did not seem to correlate strongly to pretest scores, while the Γ tended to stay the same pre/post or get worse.

(Ran out of time, much of the talk was presentation of detailed results.)

CM06: Tutorials on Coulomb's and Gauss's Laws – Zeynep Isvan, University of Pittsburgh (sponsored by Chandralehka Singh, clsingh@pitt.edu)

Five tutorials were developed to address issues with Coulomb's Law and Gauss's Law, plus superposition and symmetry. These were administered in interviews, refined, and then given after traditional instruction, with a pretest before the tutorial. The tutorials replaced some recitation activities, so the treatment group and control group had about the same amount of time on task.

Pretests tended to have low scores after traditional instruction, and the gains after tutorials were very high (0.7 to 0.9), with the best gains coming from those with the lowest pretest scores.

Tutorial groups outperformed control groups (including honors students) on the final exam, and even outperformed upper-level physics majors on the posttest.

CM07: Bridging Coursework to the "Real World" – Jeffrey Marx, McDaniel College, jmarx@mcdaniel.edu

The subjects in this study were 9th grade students. The experimental group were given discussion topics to work with, control groups were not. The EBAPS beliefs probe was used to see what students felt about science. Two important clusters of items are the "reality" cluster (which is pretty much what it sounds like), and the "source" cluster (where knowledge comes from, internal or external).

The discussions were in small groups after individual reflection on the topic. Then there was a whole-class discussion and collection of group papers written on the topic. All students had to do the papers, experiment and control.

The scores on the group papers increased over time, plateauing at a fairly good level about halfway through the term. All students followed fairly similar paths, so it's hard to find correlations.

The experimental group had improvements on the EBAPS, while the control group's scores worsened. A relatively small intervention had a noticeable positive effect on attitudes, even if it didn't affect scores.

CM08: Investigating Students' Ideas About X-Rays While Developing Innovative Teaching Materials – Spartak Kalita, Kansas State University, spartak@phys.ksu.edu

Part of the Modern Miracle Medical Machines (MMMM) Project.

Quickly went through a large amount of material about the preliminary studies and earlier phases of the project.

The current phase involved teaching interviews in which a black box containing clear and opaque Lego bricks was used as a model for X-ray mapping. Beams of light and a PASCO photodetector were used to determine where the object was, as well as something about its density. A 1-D version was also used to show how the thickness of a translucent material attenuated the beam, so that intensity and thickness could be related.

After working with the Lego model, students were asked to make statements about a representation of an actual X-ray scan, and there was some transfer. The model did seem to help in understanding the way X-rays work.

CM09: Transfer of Knowledge from Everyday Models to a Complex Situation – Bijaya Aryal, Kansas State Univesity, bijaya@phys.ksu.edu

Part of the MMMM project, specifically looking at Positron Emission Tomography. Teaching activities sought to bring in prior experience and look at appropriate and inappropriate applications thereof.

PET uses positron/electron annihilation and detects the gamma rays that result, backtracking them to figure out where the annihilation "collision" happened.

There were two main activities used to help students figure out how PET works. One had two PASCO carts behind a screen, using the spring-plungers to have them explode apart. Students needed to use the difference in time of emergence (if any) to figure out where the carts were to begin with. The other one had a series of LEDs hidden behind a milky-translucent round plastic "cake box", with LEDs lighting in pairs. Treating these bright spots as places where rays were detected in the outer ring, students would figure out where the rays all seemed to be coming from.

Once these activities had been performed, students were asked to figure out how PET could find a tumor, after being given relevant operational information. Some of the beliefs triggered in this activity included:

"Central Tendency" – The explosion always happens at the center of the cakebox, or at best at the center of the line connecting two bright spots.

"Closer is brighter/bigger" – Often inappropriately used to bolster the central tendency.

"Closer is quicker" – This is appropriate to the problem, but was rarely activated in cases where the cakebox was the first activity. However, when the carts-behind-screen activity was first, this was activated frequently.

Students often have the right ideas in their heads, but need a good cueing activity to trigger the right ones at the right times.

CM10: Tracking Eye Movements in Students' Judgments of Realistic Motion – Jose Mestre, University of Illinois at Urbana-Champaign, mestre@uiuc.edu

This study used animations of balls on tracks. One track was level, while another has a dip in the middle. The realistic case has the ball on the dipped track speed up while going down, slow while going up, and win a race against the ball on the flat track. The animations had this version, as well as various unrealistic versions (such as constant horizontal speed on the dip).

Students always picked out the correct animation when only one ball was shown at a time. However, those who had been exposed to some physics often picked an incorrect 2-ball case as the correct one (possibly misusing concepts of energy conservation). But totally naïve students correctly picked the right animation. Do the naïve and experienced students look at different things?

Eyelink II software used to track eye movements of subjects while they watched the animations. Students whose eye movements were being tracked seemed to do better at noticing the anomalies, and there were other confounding variables that still need to be teased out.

The real-time animations were almost too fast for the human eye to track properly, leading to jumpy segments called saccades. Blinking creates weird jumps as well. When the eye track went back and forth over the same area, it indicated student confusion, a "huh?" moment.

Subjects were asked to rate the "reality" of each animation, and often gave similar ratings to animations that had vastly different eye movement patterns. Eye tracking shows a lot of potential, since as an instinctual action it doesn't lie as often as self-ratings.

CM11: Research and Instructional Strategies for Student Modeling of Microscopic Friction – Edgar Corpuz, Kansas State University, eddy@phys.ksu.edu

Students start with a macroscopic model where friction increases as roughness does ($F_k = \mu N$ for those who have had formal experience). When asked to describe friction on a microscopic level, they drew diagrams showing very uneven surfaces.

Next, they were given an activity designed to challenge this model. Two metal bocks each had one side polished very smooth, and another side deliberately roughened. They were asked to predict whether the rough sides or smooth sides would have more friction, and then tried it out. The smooth sides "cold weld" (if you press them together just right, you can lift one and the other will remain sticking to it), resulting in much stronger friction than the rough sides.

An activity using crumpled and smooth paper pulled across a transparency slide followed, showing the students that contact area was important in some cases. This led via scaffolding within the Zone of Proximal Development to adding a cA term to the friction equation, and a graph of friction versus roughness that resembles a U (high friction at either very low or very high roughness).

A Karplus cycle was developed to move students through the above models. Started with sandpaper and the basic "more roughness means more friction" model, then the metal blocks to break that model, then the paper and transparency to make a new model.

CM12: And Then A Collision Occurs – Dave Van Domelen, Kansas State University, dvandom@phys.ksu.edu [yeah, a lot of KSU folks in this session]

As part of development of the Problem Decomposition Diagnostic during graduate work, it was noticed that students tended to have problems recognizing a conservation of momentum sub-problem when it appeared as part of a multiple concept problem. New work was done recently to probe that situation more deeply.

Students were offered extra credit for completing an online survey/quiz in which they were asked to identify the sub-problems in three multi-concept problems taken from the PDD, and to explain what they would do to solve each sub-problem. One problem was a deliberate red herring that had nothing to do with collisions, and another turned out to have some order-of-events issues, so its results were suspect. This left a problem in which a pendulum of mass M swings down to strike a box of mass m. The pendulum ceases forward motion and the box starts to move, eventually slowing to a stop from friction. The problem would ask students to find the coefficient of friction. An expert solution would be broken into three pieces: the downswing of the pendulum, the collision, and the subsequent motion of the box (the box and pendulum are not assumed to be the same mass).

72 students in a second semester calculus-based introductory physics course completed the survey and gave permission for their responses to be used. Of these, 31 picked a two-part solution, with most explanations omitting the collision, although a few combined the collision and the subsequent slide into a single sub-problem. 2 students explicitly invoked velocity conservation. The remaining 41 had some sort of three-part solution, although a large number of them used energy conservation for the collsion. Only 11 invoked energy conservation correctly, and one invoked velocity conservation.

8 interview subjects were chosen from this group (5 with incorrect survey answers, 3 with correct or close to correct answers). A physical version of the pendulum

problem was set up for them, and they were able to manipulate it in any way they wished. Students were audiotaped.

Common issues included:

- Remembering momentum in the first place. 4 students did not use the term at all until the interviewer brought it up near the end of the interview.
- Dichotomous elasticity. The idea that if a collision is not fully inelastic, then it's fully elastic. 2 students explicitly invoked this, while 5 students had solutions that implicitly assumed it.
- Mixing the collision and slide together. 4 students did this in some way, usually by trying to consider how much energy was lost to friction of sliding during the collision, and using the **total** distance slid as a variable for that.
- All three of the "correct" survey respondents stated that momentum was only conserved in cases where energy was conserved (2 students) or in undefined "ideal" situations (1 student). One of the other five students did not believe momentum was conserved in this problem, but gave no clear reasoning.

One subject brought up the issue of "the collision doesn't belong to either the pendulum or the box," despite that not being part of the interview protocol (although it probably should have).

No obvious suggestions for improving teaching of problem-solving yet, but the issues of conservation confusion and dichotomous elasticity should be addressed.

Labs Committee Meeting

Winter 2007 – Seattle WA

The meeting will be joint with AAAS, and using their protocols. These include a registration kiosk that's automated, and a "ready room" for speakers where you turn in an electronic copy of your talk at least 24 hours in advance (where possible). The ready room personnel will pipe your talk files to the room where your session meets, so no fiddling with plugging in laptops and the like. Papers MUST be in electronic form, and will be 10 minutes including questions as per the Syracuse meeting.

At 5:30 each night there will be poster presentations lasting 75 minutes.

Summer 2007 – Greensboro NC

This meeting will be held at a conference center several miles from campus, but the workshops will be held on campus, and dorm housing will be used. So, if you had trouble with the shuttle service at Syracuse, you may wish to arrange for either hotel rooms at/near the conference center, or rent ground transport of your own.

One proposed workshop would use data taken at the Lowe's Speedway (NASCAR track) nearby, collecting the data on site prior to the meeting. It was also suggested that some pit crew members be invited to a poster session, possibly looking to retired crew since the active crew would be kinda busy in August.

Funds are being raised to help bring North Carolina high school teachers to the workshops.

A CCLI session was proposed to help PIs demonstrate the effectiveness of their grant work, so as to help them avoid having their CCLI grants slashed.

Another proposal was to have REU students from summer 2006 attend the meeting in 2007 and present their findings (also allow summer 2007 REU students if they can get something together in time).

Winter 2008 – Baltimore MD

We should have a crackerbarrel on the use of "canned" real data sets for simulation labs. Astronomy data, accelerator runs, etc. University of Wisconsin at Whitewater is currently doing something like this, having gotten a last few runs out of their accelerator before it was shut down, and using that data for labs.

Ideas were solicited for an integrated presentation session on some topic (i.e. Gauss/Ampere Laws), covering lecture, lab, demos and so forth.

Possibility of a sponsored session touching on the creationism issue. There was significant interest in this, but it may be hard to tie something in to the Lab Committee mandate.

Wednesday, July 26, 2006

EE – Quantitative Assessment of Cognitive Models (Invited)

EE01: Uses and Limitations of Epistemological Surveys for Informing Course Design – Tim McCaskey, University of Maryland, mccaskey@physics.umd.edu

There's a difference between "public" epistemology that people display more readily, and the true "personal" epistemology that they really hold to. If we teach with an epistemological mismatch, there's conflict, so it's important to get at the personal epistemologies of students.

The MPEX2 coarsely probes epistemology, and a modified FCI was used in conjunction with it to see if the subjects' formal knowledge had been reconciled with their personal experience.

MPEX2 probes three clusters: Coherence vs. Pieces, Concepts vs. Equations, and Independence vs. Authority (source of knowledge). It uses a 5-point Likert-like scale and contextual items similar to those on the EBAPS (Epistemological Beliefs Assessment for Physical Science). The MPEX2 is not really designed to probe the beliefs of the individual student (and does poorly at this), it is meant to look at the class as a whole. They have not factor-analyzed the MPEX2 due to a model-based disagreement with how factor analysis works.

If the instructor has epistemological goals within a strong constructivist framework, and materials to support this, gains are possible in the MPEX2. Otherwise, student attitudes tend to worsen. Just having epistemological goals is not enough, however, the materials are vital.

The course under examination here was started up by Prof. Redish, who then passed it on to two other professors. Professor B used all of the materials developed by Redish, while Professor C used only the tutorials. They shared the same TA pool and lab course. Professor B had epistemological goals, but was inexperienced. Professor C was more experienced, but lacked epistemological goals.

When Redish taught the course, the class had gains in Coherence and Concepts scores, but no real change in Independence vs. Authority. B and C did about the same overall, with no real gains or losses (so both were ahead of the normal classes, at least). B had better attendance, though.

A split-task version of the FCI was given to all of the students to see if B and C had any significant differences. Students were to circle the answer that made the most sense to them, and put a box around the answer they thought a scientist might give. B's class was more likely to see students put both box and circle around the same answer. For instance, on N3L problems, 58% of B's students had the same correct answers for circle and box, while only 48% of C's did. Meanwhile, 20% of C's were unreconciled (circle and square on different options), and only 8% of B's were. (The remainder of responses had the circle and square on the same incorrect answer.)

In addition to B and C, there was an instructor coded as D, who taught the same semester in a totally traditional mode. Redish had the best reconciliation percentage of all four, and D had the worst.

The split-task FCI was found to be more sensitive than the MPEX2 or the raw FCI at detecting issues of reconciliation. The MPEX2 is simply too crude on that measure.

EE02: The Relationship of Coherence of Thought and Conceptual Change to Ability – Jim Minstrell, FACET Innovations (presented by Pamela Kraus), jimminstrell@facetinnovations.com [see their webpage for more on what a facet is...in short, it's a way of thinking about a topic, similar to a conception/misconception or a pprim].

Question: are students coherent across different kinds of motion?

Hypothesis: Yes. The majority are either correct or rely on a single bad facet.

Result: No. Fewer than half actually use a coherent way of thinking, the rest bounce around different facets.

About a thousand middle school (7th grade) students were pre/post tested on force problems involving objects that were speeding up, at a constant speed, or slowing down. Students were allowed to choose more than one response, and often there were multiple correct responses.

Speeding up: About half showed evidence of the "force follows velocity" facet (must increase F to increase v).

Constant speed: About half the class showed evidence of the "constant speed needs a force" facet.

Slowing down: About half the class showed evidence of the "Non-zero and decreasing net force needed to be slowing down" facet.

Different facets were more popular in different cases, but a consistent pattern was theoretically possible. Only about 21% of the class used all three of the facets above post-instruction, while 39% used all three pre-instruction. Many patterns emerged, however, and only about half of the students fell into "popular" patterns that were used by at least 3% of the total group. The other half used patterns that very few other people used.

The instruction in this case was a new curriculum that didn't cover all of the topics that were on the instrument.

The incidence of several incorrect patterns increased between pretest and posttest, but so did overall coherence.

[Flashed past some big comparison graphs I wasn't able to interpret. I think maybe it showed that coherent students did better than incoherent ones.]

We need to assess coherence, and it should be taught explicitly in the curriculum.

EE03: Conceptual Dynamics: Understanding Changing Student Views of Force and Motion – Ron Thornton, Tufts University, csmt@tufts.edu

[Note: Rapid motion through slides, combined with font issues that messed up the formatting of the slides, meant that my notes were a bit more scattered than I'd like for this talk.]

"Conceptual Dynamics" is a phenomenological framework used to look at how conceptual understanding changes...or doesn't.

Student domains of applicability can be different from those of an expert. Simultaneously holding contradictory or inconsistent views does not make students stupid.

It's best to talk about "student views" as a more neutral term, rather than potentially pejorative terms like misconception. A student view (SV) defines a domain of consistency, which may not be the same as that defined by a Physicist view (PV).

Use of large scale evaluation lets you see real but rare SVs, and to better track changes in SVs over time.

"View Equality" model has all SVs transform into the PV at about the same time. All starting points are roughly equal.

"View Hierarchy" model defines a chain of SVs that students evolve through before reaching a PV. i.e. SV1 becomes SV2 becomes SV3...becomes PV.

Even in a limited conceptual frame, a student can hold multiple SVs at the same time, or even SVs and PVs. This is a transitional state.

Evolution between views is gradual, not punctuated. There's no situation where the student will suddenly leap through several steps.

Students have four sets of laws of motion. The rules are different for increasing speed, decreasing speed, constant nonzero speed and zero speed. The PV is that Newton's Laws are the same everywhere and at every speed. But the SVs are scattered, context-dependent, and while the total number of common SVs is limited, a single student can hold onto several at once.

A transitional state between SVs may be better than being solidly in a particular consistent SV. This is similar to Bao's Model Analysis "mixed state", in which two or more views exist in some combination.

Students tend to reach the PV first in the decreasing speed case, then either the constant speed or increasing speed case. Force proportional to speed is a view held by only about 1/3 of the class, but those with that SV are less likely to move to a PV.

Generally, students stay still or move up through their view hierarchy. The View Equality model does not seem to hold.

EE04: Theoretical Analysis of Models and Methods in Quantitative Assessment – Lei Bao, the Ohio State University, lbao@mps.ohio-state.edu

[Note: The talk eventually moved to a series of rapid-fire equations, and I stopped taking notes at that point because I couldn't keep up. So it's longer than this summary would indicate.]

Considering the movement from Naïve to Mixed to Expert states, similar to EE03, Bao used Item Response Theory (developed by Georg Rasch in the 1960s). Variables include the discrimination factor (a), student ability (θ) and item difficulty (b).

Model-learning is seen as a transition between dichotomous states. Instruction was presented as a particle-bombardment metaphor, with different kinds of transitions having different energy levels.

Eventually a means will be developed to measure this, but for now it's just theory.

FG – PERC Bridging Session: Discipline-Based Education Research in other Disciplines (Invited)

[Note: due to weird scheduling issues, FG01 and FG03 were presented, then the remaining talks from session FI came up to use the larger room to present FI03 and FI04. FG02 was cancelled due to the speaker not being able to make it. The discussion section of FG was also cancelled.]

FG01: Use of Concept Inventories to Identify Misconceptions in Thermal Sciences – Ronald Miller, Colorado School of Mines, rlmiller@mines.edu

Chemical Engineering educators have discovered the FCI, and were inspired to try something like it for their own domain, resulting in the Thermal & Transport Concept Inventory, or TTCI. Rooted in Mikki Chi's framework of emergent processes, the TTCI covers fluid mechanics, heat transfer, thermodynamics and mass transfer.

Examining key student misconceptions, they're hoping for an instrument that is useful across several disciplines.

Development started with a Delphi study (a reasonable explanation is provided at http://en.wikipedia.org/wiki/Delphi_method) to identify the most important and least understood (by students) concepts. This was followed by open-ended interviews that were used to develop distracters in multiple choice items. It was beta tested at six engineering schools, administered in pieces as part of a validity/reliability check.

About 10 "best" topics emerged from the Delphi method, including things like energy vs. temperature, energy vs. heat, and the Ideal Gas Law.

Crosstab analysis (cross-tabulation, an example of which can be found at http://www.custominsight.com/articles/crosstab-sample.asp) was used to search for patterns among related questions to find pairs of conceptually-related answers. For instance, option e on question 3 and option d on question 5 appear together frequently, and both are incorrect, suggesting there might be a common student view that links them. Factor analysis was also performed on the items.

Persistent pairings of misconceptions/confoundings included:

- Energy vs. Temperature
- Steady State vs. Equilibrium
- Transfer rate vs. Transfer amount
- Energy quantity vs. quality (i.e. useful work, why do we have an energy shortage if energy is conserved?)

There is no such thing as an item that is too easy on this sort of thing.

One revision underway has been to add more rate vs. amount items and try the TTCI out on seniors in chemical engineering.

Context-richness was used where feasible, so that distracters might be more plausible.

Considering the consumable and transitional points in the Chi framework, how can seniors with such deep backgrounds in thermal sciences still have robust misconceptions such as confusing rate and amount? [Transcriber's aside: given how much rate vs. amount shows up in physics as well, this may well be something that's either learned very early, or even "wetwired" into us.]

FG03: Building the Biology Concept Inventory Using Ed's Tools: An Online Response Analysis System – Michael Klymkowsky, University of Colorado at Boulder, http://www.bioliteracy.net and http://www.colorado.edu/ScienceEducation

There exists some connection between assessment and pedagogy. A feedback loop would be the ideal connection, but for the most part in K12 education it's a one way street from assessment to pedagogy.

"Reflex" or "thin" assessment (i.e. just the Knowledge level in Bloom's taxonomy) leads to a simple pedagogy. Conceptual "fat" assessment leads to a more sophisticated pedagogy.

Attempting to do a concept inventory for all of biology is a rather large task, so the developers asked "What is biology really about?" to help narrow things down. [Aside: in later discussion with the speaker, I pointed out that I'd once told someone else that a BCI would be a bit too big a rock to lift...after all, there isn't a Physics Concept Inventory. He agreed, but pointed out that calling it the Biology Concept Inventory staked out important intellectual property space. Can't really argue with that. ☺]

Important questions resulting from the discussion included:

- Why are there so many types of organisms?
- Why have they changed over time?
- Why are they related? (sub-question: Why is this relationship hierarchical?)
- How does it all work? (cell, ecosystems, etc)

From this, some basic concepts were chosen to focus the BCI on, including evolutionary processes, ecological and environmental interactions (that drive evolution), cellular organization (and other detailed contexts/constraints).

So, what might conceptual problems in biology look like? Examples include, "How might mutation have a creative effect?" or "How would altering a molecule here change thus and such process?"

A multiple choice objective inventory was created with these topics in mind, and now the goal is to find ways to show student confusion (i.e. get good distracters, as well as good questions that lend themselves to good distracters). They wanted to avoid simple memorization items.

A fairly standard test development cycle was followed, using a program called "Ed's Tools" to help analyze the interviews. Simple jargon-dump items were rejected or rewritten. The BCI was anticipated to be ready for release in August 2006, so check the webpages listed above.

Ed's Tools is an online database construction program. It helps in allowing multiple coders to work on the same interview data, provides color-coded tags for potential distracters, etc. If a single line is coded for multiple things, it shows as gray.

"Reflexive" responses are a major problem, as students just parrot out memorized responses without looking outside the immediate context. Careful coding of interviews can help researchers figure out ways to break the student out of a tight context and possibly reach an Application level of understanding.

Due to vocabulary issues and the inevitable jargon of a specialized field, the BCI isn't as good at being a pretest as some might like, especially when dealing with complete novices.

[Interesting aside based on one of the BCI items. In addition to the dominant and recessive alleles most of us get taught about in high school biology, there's far more common semi-dominant alleles that allow for new traits when combined, rather than simply expressing one or the other trait. This seems to be something as basic to biology as N3L is to physics, yet I'd never run into the concept before.]

FI – Bridging Multiple Representations: From Cognitive Science to Physics Education (Invited, second half only)

FI03: Applying the Methods and Tools of PER to Reveal Isolated Sets of Knowledge in Physics Students – Mel Sabella, Chicago State University, msabella@csu.edu

Using representations to better understand where students are, basically. Since sauce for the goose is sauce for the gander, multiple representations are likely as helpful for the researcher as they are for the students. So let's borrow some from Cognitive Science.

At the moment, we seem to have the following cycle:

1. Students don't do very well.
2. We research the problem.
3. We change instruction in light of the research results.
4. GOTO 1

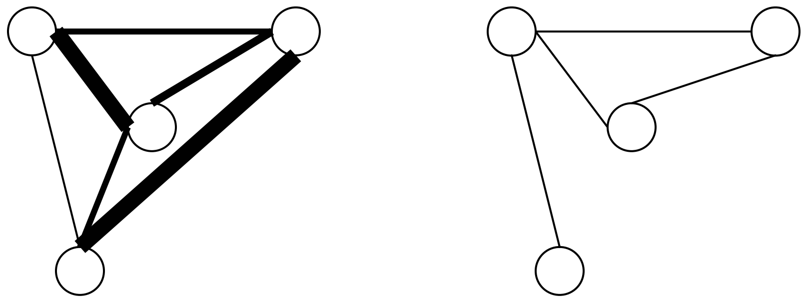
Not very heartening, is it? And it seems that no matter how many iterations we run, the cycle always goes back to point 1. Do we need a new process, or is there some other explanation?

Students often give snap responses, which are only locally coherent and connected more to intuition than to formal knowledge. It's a framing issue for them, and tends to mean that you don't actually get to probe their formal understanding. Meanwhile, experts better connect their intuition to formal knowledge, resulting in global coherence. Experts activate larger sets of knowledge than novices do, and these sets are more reliably linked to other sets, not to mention more internally consistent.

A group of novices in algebra-based courses was interviewed, using the force sled page of the FMCE. When a snap response was given, the student was asked to explain it in more depth, (slowly) eliciting their formal knowledge. When conflicts between formal knowledge and intuition came up, students attempted to form a syncretic view that incorporated both, then wobbled back and forth between the contradictory points. The more the interviewers asked for explanations, the more the students were forced to move away from intuition, displaying holes in their formal understanding.

Exam performance was found to not be a good tool for looking at formal understanding, as the timed nature of tests often resulted in the intuitive snap responses coming to the fore. It's necessary to probe deeper to get past their fragmented knowledge.

Researchers looked at the schema structure students had, the network of nodes of knowledge. They tried to find ways to activate these schema. Student schema tended to be fluid and dynamic, brought up by external cues.



On the left is a strongly linked set of schema. Everything is linked to everything else, and most of the links are robust. On the right is a weakly linked set, where none of the links is particularly strong, and there aren't direct links between all of the pieces. Each circle represents some area of knowledge.

To analyze problems better, reasoning maps are used. These are flowcharts with color-coded boxes that indicate the type of knowledge (i.e. forces, work/energy, etc) and showing how they fit together.

If knowledge is too fragmented, then a student who is cued to start in the wrong area may be unable to move to the correct concept, and be stuck so badly that no amount of extra information will get them unstuck. We need to both try to start them in the right part of the concept map, and help them link things together better so that they're less likely to get stuck.

FI04: Roles for Thought-Experiments in Model-Based Learning – John Clement, SRRU/University of Massachusetts Amherst, clement@srri.umass.edu

This talk focused on various thought experiments used to develop an explanation for the behavior of springs, as asked in the following question: Given two effectively massless springs made from the same material but coiled at different radii, which will stretch more under the same force, and why? Various models were shown as students progressed through trying to answer what is not a trivial question. Thought experiments are notoriously hard to trust.

It was desirable to partition thought experiments into some sort of useful typology. Can we find imagery indicators to help analyze them?

In analysis, they picked out various "Imagery-Related Observations" by the subject, such as "looks like", "feels loose", and any visualizations or gestures. It helped to have subjects with a lack of prior formal training on the topic, so that they wouldn't simply recall canned knowledge.

"Evaluating a model by running it" is a useful step in thought experiments, and can lead to predictions that can be evaluated. New models are devised to fix problems revealed when running an old one, and you iterate until satisfied.

An Untested Thought Experiment (UTE) predicts behavior of untested, concrete but absent systems. In other words, the system could really be tested, it just isn't available for that purpose at the moment and you haven't done so in the past.

Evaluative Gedanken Experiments (EGE) are UTEs designed or selected by a subject to help evaluate a theory, or component of a theory, such as a concept or explanatory model. EGEs can be used to confirm or disconfirm, but can also catch you on the horns of a dilemma when the model makes a counter-intuitive prediction. This is often a good thing, though.

EGEs are traditionally thought of as being mainly disconfirmatory in nature, but they need not be.

Think-aloud protocols are good for documenting thought experiments.

Multiple thought experiments can help triangulate on a more convincing result than one alone.

Both broad and narrow types of thought experiment are needed, and have been observed in classroom videos. These include elemental mental simulations, "base of analogy" thought experiments, application of explanatory models, and full EGEs.

Thursday, July 27, 2006 - PER Conference

TP-A: Issues in Concept Inventory Development

TP-A01: The Switch Effect in Quantitative Assessment: Models, Techniques and Applications – Lei Bao and Jing Wang, the Ohio State University.

A lot of work has been done recently regarding order of items and what effect it has on student performance. The gist of this poster is that in cases where items are sufficiently familiar to the student, the ordering is immaterial. In other words, their own memories of similar problems cue them sufficiently. However, when an item is sufficiently novel, then ordering effects matter.

TP-A02: Reliability, Compliance and Security of Web-Based Pre/Post Testing – Scott Bonham, Western Kentucky University

In a trial with 186 students using an online assessment system, a number of "cheating" events were looked for in the logs. These included students printing out the problems, excessive reloading, copying of responses, etc. No cheating events were found. While there are forms of cheating that they couldn't check for (such as having a tutor sitting over your shoulder), what they could see showed no cheating.

Participation rates in the online assessment were comparable to participation in the paper versions, if not slightly better. There was an overlap between students who did the paper versions and students who did the online version, but neither was a subset of the other.

Essentially, concerns such as "students will cheat on online assignments" or "students won't do the online assignments" don't seem to be significant problems, so long as you're careful.

The online assessment was developed in response to accreditation pressure to do more assessment. The results of the online version in terms of student outcomes were comparable to the results of the paper assessment.

TP-A03: Ed's Tools; Web-Based Language Analysis Tools for Concept Inventory Development – Isidoros Doxas, Kathy Garvin-Doxas and Michael Klymkowsky, University of Colorado at Boulder

Ed's Tools (see FG03) is java-based, making it flexible for implementation. It can also be used to deliver assessment instruments online, and will be used for the Biology Concept Inventory. [Most of the poster repeated stuff from FG03.]

TP-A04: Are They All Created Equal? A Comparison of Different Concept Inventory Development Methodologies – Rebecca Lindell, Elizabeth Peak and Tom Foster, Southern Illinois University at Edwardsville

A distinction is drawn between test development and instrument development, and further between concept inventories and other forms of instrument. Concept inventories should be developed with significantly more methodological rigor than one might find in a midterm exam or even an instrument like most Physics content surveys currently in existence

This poster rated several instruments on the basis of concept domain (the basis of distracters), test specifications, field testing, item statistics (such as difficulty and discrimination), validity (criterion, construct and content) and reliability. Using published information on the instruments (which was scanty in some cases), the following instruments were rated (acronyms expanded where I know them and they're not in the glossary already): ADT, BEMA, CSEM, DEEM, DIRECT, EMCS, FCI, FMCE, LPCI (Lunar Phases Concept Inventory), TUG-K, WCI. It was concluded that a lot of instruments that called themselves concept inventories did not fit the criteria for such. In fact, only the LPCI met all of the criteria.

The steps involved in instrument development methodologies are:

1. Identify the purpose of the instrument
2. Determine the concept domain
3. Prepare the test specifications (what does it need to do?)
4. Construct an initial pool of items
5. Have the items reviewed, and revise as necessary
6. Hold preliminary field testing, and revise as necessary
7. Field test a large sample that is representative of the population
8. Determine the statistical properties of the item scores, and eliminate bad items
9. Design and conduct reliability and validity studies
10. Develop guidelines for administration, scoring and the interpretation of test scores

TP-A05: Validation Studies of the Colorado Physics Problem-Solving Survey – W.K. Adams and C.E. Wieman, University of Colorado.

The CPPSS is an attempt to develop a problem-solving survey without the use of physics concepts, in order to avoid situations where a student has simply learned the correct answer by rote memorization, as well as situations where the student is unable to solve the problem due to a lack of content knowledge. It looks at a general skillset, and found that skills don't form a "tree" where you have to acquire certain skills before others. The skills seemed to show up in just about every order and combination.

One sample item was the "Pyramid of Giza" problem, in which students have to estimate how many blocks there are in the pyramid, and from that how many workers are needed. Another is a base problem from the Jasper Woodbury series (located at http://peabody.vanderbilt.edu/projects/funded/jasper) of video story problems. In the case of the Giza problem, students were given details of how the pyramid might have been built, so that there wasn't room for alternatives such as concrete-pouring or alien builders. This constrains the set of possible solutions a little better, making it easier to evaluate the results.

The survey is still in development, but the plan is for it to be partially multiple choice and partly open response.

Discussion Session: [Where I knew the speaker, I have noted it. There are a couple of statements that may have been made by a known speaker, but my notes were not clear enough.]

Maloney: Be very careful not to use the term "inventory" too loosely, as it has a very specific meaning in education circles. It would be akin to using energy and power interchangeably.

Lindell: Many "concept inventories" are not really inventories. "Instrument" is a better general term. An inventory is not a self-report or a survey. It looks at particular content structure, needs at least five items on each aspect or issue within the content (while surveys are often content with 1-2 items per aspect).

Maloney: The FCI is closed to being an inventory, the FMCE is an inventory despite the lack of the word in the name. The CSEM isn't even close.

Unknown (possibly Maloney): We should not ignore the way that 40+ years of education research has defined things, especially if we want to work with people in other education research areas. Clearer nomenclature makes it easier to use the right instrument for the job.

Klymkowsky(?): Perhaps the BCI will be renamed an Instrument, as at the moment a rougher grain with fewer items is desired. Perhaps the successor instrument will be more correctly called an Inventory.

Unknown: "Instrument" is a loaded term for physicists, leading us down the wrong path. Perhaps "evaluation" would be a better class-generic term.

Invited Talks

IT01: The Effects of Inquiry-Based Instruction on Elementary Teaching Majors' Chemistry Content Knowledge, And Their Views About Teaching Science – Michael Sanger, Middle Tennessee State University (previously University of Northern Iowa), mjsanger@mtsu.edu.

K12 standards advocate teaching by inquiry, and teachers tend to teach as they were taught, both pre-service and in-service. Ergo, it'd be a good idea to teach by inquiry to teachers.

However, as in physics, many chemistry instructors view inquiry-based courses as watered-down, and resist this. It's necessary to have non-anecdotal evidence to convince them that inquiry-based courses are not watered-down.

Two groups were looked at in this study. The freshman General Chemistry (GC) students were the control group, while elementary education students in a freshman Physical Science (PS) course were the experimental group. The same instructor taught both courses. GC was taught in a traditional "not watered down" mode, while PS was taught using University of Northern Iowa's PRISMS materials.

Topic 1: the molecular definition of density. Asked to define density without using the words "mass" or "volume". Many simply used synonyms. The success rates were 44% for PS and 22% for GC, with an p-value of about 0.04.

Topic 2: compressibility of solids, liquids and gases. Students were asked why gas was more easily compressed than solids or liquids. All PS students were able to answer this successfully, and 89% of GC students ($z=1.89$, so not quite significant by the usual standards).

Topic 3: structure of molecules. Students were asked to draw the structures of NaCl and other molecules, and describe the forces involved. Similar results for both courses. PS did better on diagrams and GC did better on forces, but not to any statistical significance.

In the next phase, elementary education majors in PS were compared to upperclass secondary teaching majors (UT) who had a traditional science background, and the question was asked, "Do the educators develop different views, after inquiry instruction, of how science works?"

PS liked inquiry and uses sources as supplements. The UTs decried the lack of transmissionist teaching and the lack of facts in the books. The UTs wanted a lecture component.

Some comparisons between UT and PS:

- Teaching by telling vs. learning by messing around in the dark
- "errors are bad" vs. "errors are part of the process"
- "the real world gets in the way" vs. "the real world **is** the way"

It turned out that the inquiry methods taught the actual material as well as or better than traditional methods, and the PS students had much better attitudes towards science.

IT02: Rasch Analysis and the Geoscience Concept Inventory – Julie Libarkin, Michigan State University, libarkin@ohio.edu

The theoretical perspective here is a mix of "Synthesis of prior knowledge" (Posner, Vosniadou), "Recategorization of concepts" (Mikki Chi), and "Fragments are organized into structures" (diSessa).

Example: Students commonly synthesize plate tectonics with the idea of political boundaries to get the "states as plates" result, or the idea that plates contain whole continents rather than sometimes dividing continents into pieces.

Example: Students have very fragmented ideas about magnetism in general, which can lead to difficulty in describing geomagnetism.

Example: Students equate "hot" with "equator" and "volcano" with "hot," so many come to the conclusion that volcanoes are clustered at the equator, and that the equator is a plate boundary.

These examples and others have prompted the development of a quantitative inventory, in part due to conversations with Noah Finklestein. The GCI is based on the FCI and theoretical frameworks such as grounded theory, scaled development theory and item response theory. Grounded theory in particular generated many items and found questions to ask that the researchers wouldn't have thought of on their own.

The items were then tested for validity and reliability, using 75 interviews and about 1000 subjects in a questionnaire format. There was a good spread of external institutions, and high demographic richness. The construct validity was found to be very good.

An example of a distracter found during development was that students tend to conflate gravity and magnetism, leading to claims that a wooden satellite wouldn't stay in orbit.

There were five items that were found to have significantly different results when you changed the population (male vs. female, white vs. nonwhite). These items have not been removed from the instrument, but do come with warning flags for test administrators.

The reliability has been found to be about 0.56.

Item Response Theory is used to adjust the score to reflect different levels of knowing (i.e. half vs. all).

Rasch Analysis finds which items are hard, easy and in between, then lets you scale the scores accordingly. The GCI was found to have no trivially easy items.

Items of the same Rasch difficulty are statistically similar. This, in theory, can let you use a large bank of items to create multiple subtests, as items with the same Rasch difficulty can be substituted for each other.

When your raw score is at an extreme of the range, it's harder to learn more than you already know. At the low end, you're lost, while at the high end, there's not much left to know within the domain.

GCI results show that low-pretest score students seem to learn the most, regardless of teaching style. This would suggest that even a low GCI score is not "low" for purposes of rate of learning, and ties in with the "no really easy items" result.

It should be noted that what geoscience education considers inquiry-based teaching is not really the same as what physics education does, in part because of the difficulty of performing hands-on tectonic experiments and the like.

The GCI can be found at http://newton.bhsu.edu/eps/gci.html

They are developing more items for the GCI, and trying to link up with other inventories.

When asking for demographic information during GCI development, these questions were asked at the end, in an attempt to avoid "stereotype threat" (in which reminding someone that they are part of a group thought to do poorly at something will actually cause performance to decrease).

IT03: Astronomy Concept Inventories: Foundations and Frontiers – Beth Hufnagel, Anne Arundel Community College, brhufnagel@aacc.edu

The main focus of this talk was the development of the Astronomy Diagnostic Test, or ADT. By the definitions of TP-A04, it is a survey and not a concept inventory or diagnostic.

Astronomy Education Research is not simply a subset of Physics Education Research, but as they don't have their own conference yet, AER will continue to be presented at AAPT meetings, even when the theme isn't cross-discipline work.

While the rote memorization of facts is not considered to be sufficient by educators, some is necessary in order for the meaningful questions to be answered. You can only simplify and remove context so much before you lose touch with reality.

You need to start with a good research team...you simply can't even attempt a concept inventory as a solo act. It's best to have a cross-discipline team, too. Not just across content areas, but also experts in psychology, cognitive science, statistics, etc. It took a fairly large team at Maryland two years to build the ADT.

You also need a theoretical foundation, and it helps to outline one explicitly before you start rather than figuring out what you've been doing after the fact. It's also a good idea to know what content is important before you start, rather than just assuming that whatever's being taught right now is good enough.

Introductory Astronomy instructors have rarely taken a version of the course they're teaching, having started farther up the curricular ladder and not looking back until they'd reached the point where they had to start teaching it. Astronomy PhDs may not even know how to operate a physical telescope, having done all their work with big computer controlled telescopes operated by technicians. This works about as well as you might expect when they end up in the classroom.

Astronomy's culture is strongly anti-teaching, "worse than physics!"

The ADT is aimed mainly at the sorts of things you'd expect to be covered in high school or even junior high physical science classes. There was a surprisingly strong gender gap in the results, with men scoring as much as 10-12 points higher than women. In none of the sample classrooms did a woman have the high score, and in every classroom the average score for men was higher than the average score for women. Looking at the gap by topic, it seems that cosmology items accounted for a significant chunk of the gender gap.

When looking at self-rated confidence (i.e. students were asked to rate how confident they were in their answers), men were slightly more confident than women, but no one was terribly confident. It's possible that the stereotype threat kicked in, especially on more mathematically slanted items. The "Barbie Effect" in action.

In an unrelated study, students were specifically told there was a gender effect, but not who benefited. Men told about the existence of a stereotype threat did better than men who were not told anything about it, and the performance of women suffered.

The ADT, however, did not show evidence of a stereotype threat when they tried the same thing. It's possible that the students don't care enough about astronomy to be threatened by the stereotype.

Moving on, it was pointed out that astronomy is less reductive and more integrative than physics, hence requiring a somewhat different style of thinking.

Astronomy professors fear concept inventories and similar surveys, as they worry it will prove that they can't teach. Thus, there is resistance to implementing surveys. However, playing to the student expectation of an increasing use of computer-based educational technology allows reformers to partially bypass the professors. The professors can stick with the "big picture" stuff they prefer, while the reformers implement nitty-gritty changes in the edutech. It's also possible that encouragement of collaborative learning will reduce the gender gap.

IT04: Panel Discussion.

[Comments will be coded as follows: C – Chemistry (Sanger), G – Geoscience (Libarkin), A – Astronomy (Hufnagel).]

Q: What are key parts of your area at the college level?

C: It's still somewhat diffuse, but there's a strong realization that ability at math does not equate to ability with concepts.

G: Not a lot of work being done at the college level, most of it is aimed at K12. Only about 5% of presentations at their meetings even have data.

A: Still trying to make a true inventory along the lines of the FMCE. A light and spectroscopy concept inventory might be possible. There's a lack of agreement on introductory astronomy content, with publishers driving curriculum rather than the other way around. There's very little government-funded curriculum work being done, it's all funded through the publishers.

Q: Can you get tenure as an education researcher in your field?

G: There was a recent National Academy of the Sciences workshop on this. So far, 9 geoscience education researchers have gotten tenure...it's not easy, but it's possible. GSER is still mainly populated by "switch in after tenure" types as it was in early PER.

A: Maryland offered a joint Astronomy/Education appointment, but it looked to be a recipe for failure in the tenure hunt. She left for a community college, where teaching is more valued, and still got flack over reform approach. But she did get tenure.

C: ChemEd Research is pretty well established right now, about where PER is in terms of PhDs with specific degrees in it. In some ways, CER is ahead of PER, and there's sufficient research outlets and acceptance. However, it's still best to not be the soloist at a university, as solo CER people tend to get saddled with curriculum development and left no time to actually research.

Q: Have gender studies looked at single-gender schools? Is there still a gap when comparing women from all-girl schools to the general population?

A: Women from single-gender schools did do better on the ADT.

G: No change in the gender gap with the GSCI.

Roundtable RT-A: Cross-Pollination in Science Education Research

Organizers: Eleanor Sayre (University of Maine) and Rebecca Lindell (SIUE)

Results will be available at http://www.zaposa.com/PER/cross-pollination

We broke into three groups for initial discussions:

1. Community building and finding collaborations
2. Topics across disciplines
3. Cross-training out of our own field (I was in this group)

Various comments that came up in group 3 discussion:

Tufts is running a program that brings in people with existing degrees or work-obtained skills and trains them in education.

Credentialing issues affect out-of-field training. People may need out-of-field training to succeed in teaching, but aren't required to get it by the credentialing agency, so they don't do it.

We need specialized content courses for teachers, even for those who already have degrees-in-field. Especially for them, in fact (see, for instance, the introductory astronomy teacher issue from IT03 above).

The trend towards broad certification means that out-of-field courses are even more important...that one physics course they take may be their only exposure to the subject before having to teach it! We need to make sure that teacher-training physics courses model best practices and have faculty backing.

Science education courses tend to be "one size fits all disciplines" in nature. This makes them more focused on serving the more popular disciplines, but on the plus side does expose all potential science teachers to all fields.

Physics education courses are often taught at the graduate level as seminars, which makes them effectively invisible in the line schedule (especially at schools where they're simply listed as a "special topics seminar" due to difficulty in getting new courses into the catalog).

Once out in the field, teachers express the wish that they had learned more procedural stuff, but research indicates that they simply didn't absorb the procedural training that they did get, seeing it as unimportant at the time.

We need elementary teachers to not be science-phobic, especially since more of the NCLB-mandated materials do not require science in K-3. If they neither want to teach science nor have to teach science, they're not going to teach science.

We need either good teams of content and pedagogy experts, or a single teacher who can straddle both. Essentially, we need a "Reese's Peanut Butter Cup" of content and pedagogy. And to respect both the chocolate and the peanut butter. City University of New York reportedly has had success with such teams.

Brigham Young University has turned out more physics teachers each year recently than many entire states have generated.

Report from Table 1 (Collaborations)

When teaching across disciplines, you need to ask the following questions:

1. Who would benefit?
2. What is our toolbox?
3. How would you build a collaboration to fill the toolbox?

In the toolbox are such diverse elements as:

- Topics including math, cognitive science, educational leadership and literacy
- Respect across disciplines
- Joint meetings to keep collaborations alive (such as AERA)
- Looking for people in other disciplines willing to share, listen and talk.

Report from Table 2 (Topics across disciplines)

The question was raised, how is problem-solving different across disciplines?

Consider overlap of content, like "What does it mean to know ______ for a physicist? For a biologist?"

Try to know what the key topics are in other fields (i.e. in Biology, evolution and the mechanics of cellular operation).

Do we want to look outside science, math and engineering fields, such as seeing what's done in music education?

University of Colorado is currently a very good example of cross-pollination.

[Roundtable was sort of falling apart near the end, as people had to leave to catch flights. I didn't attempt to go to anything in the final session, since I would have had to leave partway through for the same reason myself.]

And so ends another installment of my AAPT meeting notes. I hope to attend the Seattle meeting in January 2007, since it's not overlapping with classes this time.

Dave Van Domelen
dvandom@phys.ksu.edu