

Assessment and Evaluation Primer
VaNTH ERC
August 9, 2001
©Copyright Vanderbilt University 2001, All Rights Reserved

I. Purpose of this document

This document is designed to provide basic information to domain experts for the design of evaluations for modules, courses and other groupings of instruction for the VaNTH ERC. The focus of this document is on bioengineering knowledge assessment. The development of appropriate assessment tools for knowledge assessment in bioengineering is the primary responsibility of the domain project leader since these assessments are heavily based on domain knowledge content. A&E and LS experts can help domain investigators review and improve their assessment questions and procedures, but cannot provide such domain intensive materials.

II. Overall Plan

This document is a plan of action to bring knowledge –based assessment into action for the academic year 2001-2002. Our goal is to obtain data on the value of the HPL-VaNTH methodology in our target domains for presentation at the May 2002 summative NSF site visit. The plan and milestones are presented below:

1. July 2001: Distribution of knowledge assessment primer and plan to domain project leaders. (Deadline: August 10, 2001).
2. August 2001: Identification of module and/or course test beds throughout VaNTH for knowledge assessment during AY 2001-2002. (Deadline August 20, 2001).
3. August 2001: Development of testing designs (modules, courses, controls, cognitive labs, etc.). (Deadline September 1, 2001).
4. August-September 2001: Development of knowledge-based assessment questions followed by review by A&E experts.(Deadline September 15, 2001).
5. September-April, 2001: Performance of evaluations as part of module-course efforts during the academic year (Deadline May 1, 2002).

III. Assessment Methods

At the Evanston July 2001 meeting Jim Pelligrino showed the following outline of assessment needed for VaNTH (Figure1):

ASPECTS OF ASSESSMENT IN THE “HOW PEOPLE LEARN” MODEL

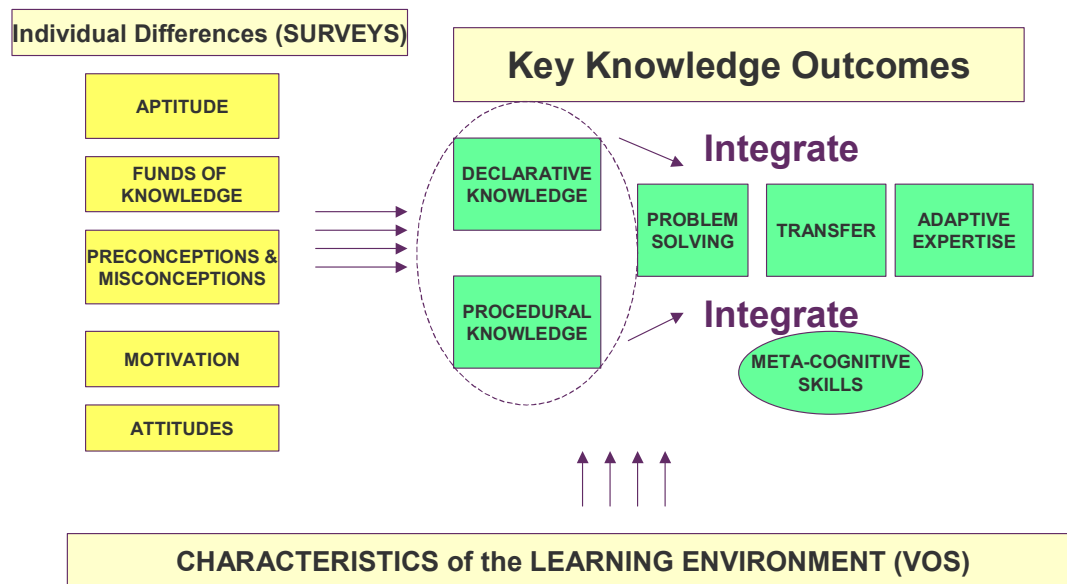


Figure 1: Assessment of HPL system.

It should be noted that methods for the assessments labeled “Individual Differences” and “Characteristics of the Learning Environment” have been developed and tested last spring. These are ready for implementation in 2001-2002. Early testing of Key Knowledge Outcomes tended to concentrate on declarative and procedural knowledge and may not have integrated problem solving, transfer, adaptive expertise and metacognition in ways that really measure the power of the HPL system. These are the assessments that are now needed for combination with the existing surveys and observations to provide a complete set of useful assessments. **It is the purpose of this document to provide help in formulating these assessments questions that could be used as part of regular instruction, or could be used in a special “cognitive laboratory”.**

IV: Examples of Knowledge Outcome Questions

We need to design a range of assessment items to target specific objectives related to a unit of study in a course, an entire course, or the overall program of undergraduate bioengineering study. A good test should objectively, accurately and consistently measure indicators of students' "understanding". In addition, a good test needs to be graded in a realistic time frame. Therefore, quite often the most common assessment methods used are multi-step word problems with partial credit for each step, multiple choice, short answer or matching activities to capture students' knowledge. These methods work well for targeting *important* concepts, and concepts students need to be *familiar* with relative to the learning objectives of the course (Wiggins & McTighe, 1998) and should be used as part of our usual course assessments. However, many traditional assessments are not sensitive to the changes in student learning expected with VaNTH's instructional methods (based on the HPL framework). Assessing students' understanding of *fundamental* concepts of the course materials, and their progress toward becoming excellent bioengineers, requires an assessment where students must integrate multiple ideas together. Therefore, students need to be able to demonstrate their ability to handle every phase of solving a complex engineering problem (e.g. problem identification, exploration of possible solutions, discovery and execution of a plan and reflection on the solution for accuracy). We should be able to refine some of our existing questions, and design new questions, to help deepen our assessment of course materials and our overall program evaluation. The following provides a short description for how we might accomplish this goal.

John Bransford discusses the nature of assessments that truly probe HPL principles in the attached document on "Transfer Appropriate Processing" (Appendix A). The concept is introduced that the amount or "preprocessing" of information for students should be reduced in order to test their ability to integrate. Otherwise, the problem is reduced to a kind of multiple-choice examination, which reduces the chance of student error on objectives not immediately relevant to the specific learning objectives for the course. For example, a question can be worded to help normalize the class to the specific content taught in the current course and less on prior knowledge students bring to the course. However, we need to make a transition toward evaluating students' ability to define problems from context, generate methods to represent the problems and ask questions about what more information is necessary to discover a solution to the problem. **In short, we need to evaluate students' ability to *qualitatively* explain a problem that leads to discovering a *quantitative* solution to a problem and accurately calculating that solution.**

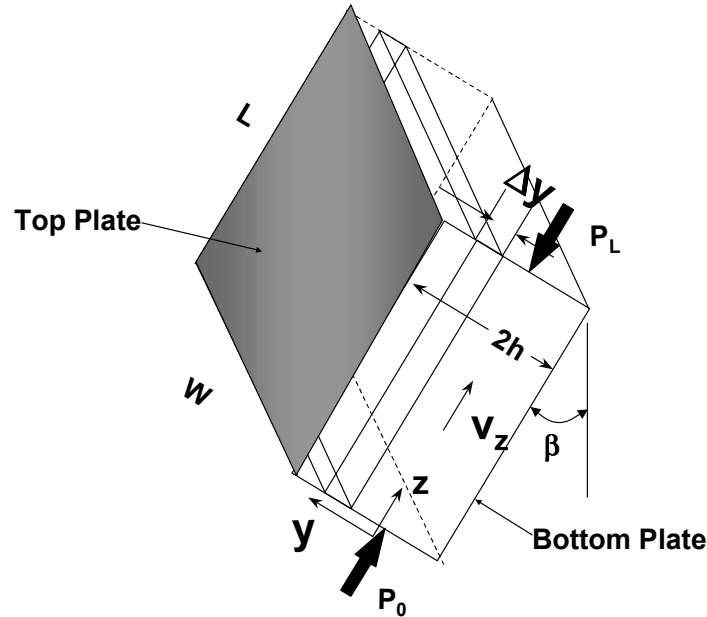
Consider an example from transport phenomena that is aimed at summative evaluation of students' abilities to apply the momentum balance to a simple one-dimensional flow field. This is shown in Figure 2. A more integrative approach might be that shown in Figure 3. While the demand for integration is only slightly greater in Figure 3 than Figure 2, it does probe deeper skills (and is a much harder problem). Neither problem could be approached without instruction and formative assessment, but Figure 3 might be more effective as an HPL knowledge-based assessment.

What has happened to the assessment question 2 as a result of rewriting the three steps into a single problem statements? One change from Figure 2 to 3 includes the removal of the categorization as a momentum balance problem. Some instructors are finding that students will perform well on multi-step problems like figure 2, because selection of the appropriate equations is not part of that problem's process. They've been changing their instruction to include discussion of how the underlying principles relative to the course relate to the specific problem. In their assessments they expect students to be able to do the same. However, even asking students to identify the underlying principles can still be a trivial move for students if momentum balance equations are the current unit of study for a course. There students should be prepared to justify why these principles apply for the specific context. Another changes from figure 2 to 3 relates to ability to formulate the equation in terms of velocity v (Figure 2c), where the question statement references explicitly the goal of deriving an equation in terms of velocity. Students are not asked why velocity is a critical feature for this class of problems; therefore, we are not measuring their ability to make these decisions. In the transformation to Figure 3 the students are given the assumption "that a single component velocity dominates flow in this system". Part of the grading for Figure 3 might include seeing if students recognize the need to write their equations in terms of the velocity. The rewriting of the fluid transport problem moves beyond students' ability to manipulate equations toward a problem that also includes their abilities to qualitatively analyze the problem based on its underlying principles.

Problems like these can be taken even further to target the development of students' problem solving abilities. Figure 4 defines an even greater reduction in support given to the students (Roselli, 2001) by elaborating on a specific context where this knowledge of momentum transport is required. There are many steps involved in transforming an open-ended question like Figure 4 to the stage of the problem solving process for Figure 3. Experts constrain the problem by making assumptions about how the fluid is flowing, what factors are invariant in the problem (Biswas et al) and how to represent the problem using an element analysis. We need to evaluate students understanding of how to manage the complexity of open-ended problems like these. Several of these kinds of problems should be offered in a course to evaluate students' progress for solving these problem across the timeline of the course and their course of study. Students need to understand that these opportunities to attempt the these complex problem is preparing them for their future careers; therefore, they should use these experience to reflect on their progress.

Figure 2: Current Declarative and Procedural Based Problem

Problem 2: Consider that an incompressible, Newtonian fluid is flowing through a slit (fluid is confined between two stationary solid plates) under a pressure gradient. This system is oriented at an angle β with the vertical axis. A sketch is shown below:



Consider that the pressure acting across the lower cross sectional flow area is P_0 and the pressure acting across the upper cross section of flow area is P_L . The overall thickness of the slit is $2h$. Consider that the axis is oriented so that there is one component of velocity, v_z .

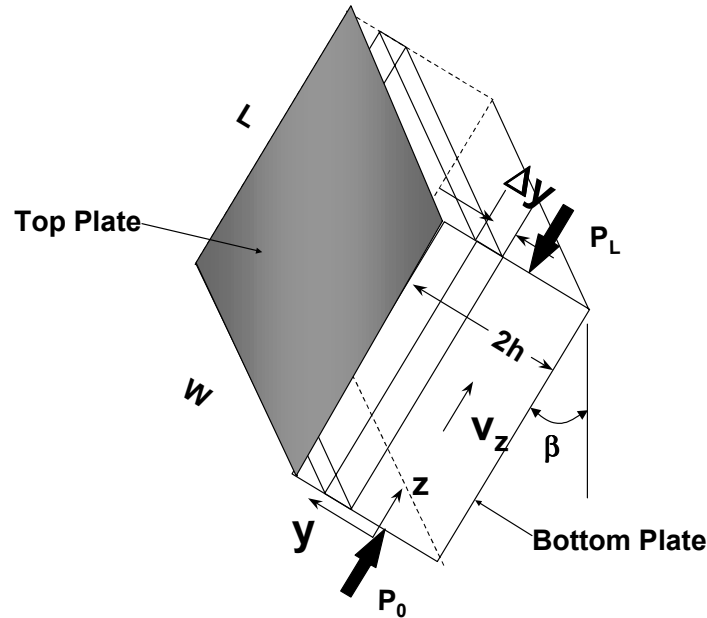
(15%) a. Write a momentum balance over the small element that is Δy in height, L units long and W units wide. Express the balance in terms of the system variables shown in the sketch and other important terms. Define any terms that you use.

(5%) b. Reduce the momentum balance to a differential equation.

(15%) c. Convert the differential equation to a differential equation in velocity v_z and position within the fluid. State the boundary conditions for the solution of the equation. Justify your choice. DO NOT SOLVE THE EQUATION.

Figure 3: Integrative Problem

Problem 2: Consider that an incompressible, Newtonian fluid is flowing through a slit (fluid is confined between two stationary solid plates) under a pressure gradient. This system is oriented at an angle β with the vertical axis. A sketch is shown below:



Consider that the pressure acting across the lower cross sectional flow area is P_0 and the pressure acting across the upper cross section of flow area is P_L . The overall thickness of the slit is $2h$. Assume that a single component velocity dominates flow in this system. This component is part of a micro-flow system in which it will be important to know the effects of orientation (angle β on the volumetric flow through the slit. Derive an equation that relates volumetric flow to system parameters, pressures and angle β

Figure 4. Blood Oxygenator

Problem 2. During open-heart surgery a patient is placed on an extracorporeal blood oxygenator. Venous blood is pumped through the device, which consists of many parallel channels that are separated by thin, rigid, gas-permeable membranes. The top and bottom membranes of each blood channel are in contact with an oxygen-rich gas. If the pump at the inlet of the oxygenator provides a constant inlet pressure, how will blood flow rate through the device change as the outlet end of the device is tilted upward by an angle β ? (Roselli, 2001).

V. Evaluating existing assessment Items and Defining New Assessments

One of our immediate goals is to define similar assessments to those presented in figures 3 and 4 that target the new learning objectives imbedded in the challenge based approach to instruction VaNTH is adopting. In addition, we need to identify difficulties students still have at the end of a course. Therefore, John Bransford suggested at the July 28th Learning Forum that the domain experts review past final exams for questions that students still had trouble solving. The Learning Scientists can work with the domain experts to explore how these difficulties relate to current theories of how people learn. Together the LS and domain experts can identify potential instructional methods to target these deficiencies.

The other critical step for domain experts is to create assessment items that target more of the higher level engineering skills (core concepts) students should demonstrate by graduation. As a starting point we suggest reviewing current test items used in a course as candidates for revision. One heuristic for selecting current test items for revision would be problems that encourage students to evaluate the problem at multiple levels, rather than “problems that are simple applications of definitions (e.g. “compute the moment of inertia of...”) or procedures (e.g. most kinematics problems) “(Leonard et al, 1996). Good problems might request students to articulate what the underlying principles that relate to the problem, a justification for why these principles apply and a procedure explaining how these principles would be applied to define a solution to the problem. This qualitative portion of problem solving is one of the dimensions of learning we are trying to engage students with the challenge based approach to instruction. The use of these new assessment items will be instrumental in determining our success with this goal.

VI: Design Methods for Evaluation of VaNTH Materials

The appendix B contains a discussion by David Cordray on the relative power of different designs for the implementation of innovations and assessments. Please review these. The following is a general guide to their use. Note that even a weak assessment is better than no assessment at all.

a. Single group vs multiple group

In many cases we will be faced with a single group that we would like to assess. David offers three ways to do this: 1. Performance threshold; 2. Norm-based; and 3. Before-after. The simplest and weakest is threshold. Here you are simply observing that this year’s students being taught with a VaNTH innovation actually do better on your tests than last year’s students. David offers some advice about improving the power of this approach, including “norming” the assessment (approach 2). The best of these single group approaches is the pre-test post-test method. This is the best and fairly simple way to measure an innovation in a single group. Its power is aided if your knowledge-based questions are well designed. If you can only study a single group, try this approach with assessment questions reviewed by A&E.

b. Multiple Groups

These are the most powerful designs as discussed by David. Please review and see if you can identify situations where a control group is feasible. We have requested a list of potential test beds from all sites. If we receive these data, we should be able to identify some cross-institutional venues for control and intervention groups.

Appendix A
**Aligning Challenge-based Learning with Assessment:
The Concept of “Transfer Appropriate Processing”**

John Bransford, July 20, 2001

Memo to VaNTH group

There is a concept in the cognitive literature called “transfer appropriate processing” (TAP) that is based on data showing that a particular type of learning activity may look good or poor depending on the measures used to assess the learning (Morris, Bransford & Franks, 1978). The idea of “transfer appropriate processing” is that we must design learning activities that fit the kinds of tasks we want people to do at the end of our courses and programs. This means that the tasks (assessments) we choose are crucial for making inferences about the quality of what was learned.

The TAP concept is relevant to BioEngineering because many ways to assess learning will fail to capture the value-added of having students engage in challenge - based inquiry that is designed to help them become adaptive experts. I think we are missing the value of many of the teaching innovations in VaNTH because our assessments are not truly aligned with what we hope our students learned. I try to elaborate this point below.

Example I: Michael’s Use of Different Measures: One paper I thought had been sent around last year but evidently got lost in transit is one that Ann Michael headed with a group of us at Vanderbilt. She taught an experimental and controls class in the diagnosis and treatment of delays in early language development. One class received a traditional organization of content where different topics covered behaviorist approaches to language therapy, another covered social linguistic approaches, another covered Vygotskian approaches, etc.

A second class of students received an early precursor of a “challenge based approach”. The course was anchored around an initial video of a therapist doing speech therapy with a child. Students generated their initial thoughts about what they noticed. Then they received multiple perspectives from a behaviorist, social linguistic, Vygotskian, etc. These perspectives were each related to the anchoring challenge video. Students then mapped these experts’ perspectives into the text.

As expected, only some kinds of assessments showed the value of the challenge-based teaching compared to the traditional. Purely factual questions were expected to be the same for both groups--and they were. Differences occurred when students received new cases--one via video and one in text format--and were asked to comment on them from the three theoretical perspectives. Students in the experimental group did a much better job on these tasks than did those in the control group.

Note that Michael’s transfer task was open-ended. She simply showed a video (or in the second case a verbal) description of an interaction between a child and a language therapist. She did NOT stop the video and say things like “how does this particular action by the therapist relate to the behaviorist concept of X or Y”, etc. And she did not

say “imagine that the behaviorist stated his diagnosis as X---do you agree? The latter kinds of specific questions move the transfer tasks from (a) open ended tasks to where STUDENTS must do the noticing and searching for relevant information on their own to (b) tasks where specific questions are heavily prompted. In short, the task now becomes much more like a multiple-choice test.

Note that “real” experts have to do their own noticing of relevant information--they can't rely on someone always prompting them on what to look for and how to define each step of their task. So I conjecture that unprompted assessments are more “authentic” in terms of what we want our students to learn to be able to do. Remember the video of the houseboat expert? To be a true expert, he needs to notice features on his own (which he did). The “analyze the houseboat task” could be made progressively easy and even trivial if we stopped the video at various points with specific questions like “Gibson houseboats come with these kinds of rails (show picture). How do these compare with the railings in the video? To anticipate the argument below, I think that many of our Bio Engineering assessments are closer to the latter (heavily prompted) than the former (more open ended with more room to assess the kinds of processes related to “adaptive expertise”. (See the discussions of adaptive expertise that have been circulating).

Example II: Research on Complex Problem Solving in the Context of Jasper: As a second example of differences between open ended and heavily prompted assessment environments, consider research on the LTC's Jasper adventures. Each adventure is an approximately 15 minute story that ends in a challenge, which requires a great deal of complex thinking and mathematical calculation. All the data needed to solve the problems have been embedded in the story line--but there is also lots of irrelevant data. Students have to formulate a plan, devise learning goals for finding the relevant information, and then working to solve complex problems.

In one set of studies we took the highest scoring sixth grade students on standardized math tests and asked them to solve a Jasper without any prior instruction on it. They were terrible at it --mainly because they didn't know how to conceptualize the overall problems, break it into steps, search for relevant information and then “do the math”. Non-mathematical college students at Vandy were also quite terrible at this--although they were a little better than the sixth graders (remember that the latter were very high achieving given standard measures of math).

A second group of students saw the same Jasper adventures and were asked to solve it. However, we structured our questions by asking a series of short sub-problems that were needed to solve the overall adventure. Once we defined these sub problems for them they did MUCH better than the previous group. They knew “the math”, but they could not define these sub problems on their own. After training in Jasper, there were large increases in students' abilities to do “unprompted problem solving” for complex tasks. (See CTGV, 1997).

Analogous issues in Bioengineering: Many problem solving tests that I have seen used in Bio engineering seem closer to the side of being “heavily prompted” rather than open ended. This is our (my) fault for not pointing out this issue sooner. It’s is one of those things that is so taken for granted by our LS research team that we forgot to mention it. (It’s a good example of the need for better “pedagogical content knowledge” on our part). The issue of TAP surfaced when we had the chance to analyze actual BioEngineering tests that were being used.(Many thanks to those who shared these tests).

A sample test question in Bio Mechanics can (hopefully) serve to clarify the relevance of TAP issues for our work. On one test item, students are introduced to a woman who has had a hip replacement that is very sore, so she has to walk on a cane. The initial problem statement includes all kinds of specific data like her weight, the length of the cane, very helpful visuals representations of the hip and angles of stress. But so much is specified that the problem essentially asks students to “do the math” without asking them to first frame the problem, decide what they further need to know, etc. These “highly specified” problems seem more analogous to the Jasper studies where we specified sub problems for students and they did quite well. But I think we can conjecture that too much specification misses the opportunity to measure important features of the concept of “adaptive expertise” --where people must start from scratch to define issues and decide what they further need to know.

How might we transform the question about the woman with the sore hip? Here’s one possibility (I’m sure this needs to be tweaked because I need a better specification of exactly what we want students to know and be able to do at this point).

A physician asks a bioengineer to help alleviate a woman’s hip pain following hip surgery on her left hip. The woman is unable to put any pressure on one of her feet because it hurts to much. He finds similar problems with all his hip replacement patients (the problem typically lasts for about one month or so). He wants ideas about how to help his patients get around comfortably in a way that minimizes their pain.

As you approach this task, briefly write down your initial conjectures about the pain and why it occurs; additional data you would like to know, and additional questions you would like to ask.

- Some of my thoughts as a novice: As a novice (my fate in life it seems), a number of questions come to mind. For example, a cane or crutches might help. Certain kinds of shoes might help as well. Maybe the pain is the result of a poor method for hip replacement. I need to understand lots of things in order to proceed. To go through this process, I’ll use the IDEAL framework sketched in my thought paper on adaptive expertise (sent to the group a few days ago).

I - INVEST the Time to see Problems as Opportunities for Learning:

First, as a student I need to understand why it is useful to INVEST my time exploring this kind of open ended problem. If I don't understand the value (I.e. This is going to prepare you for the kinds of things I'll need to do to shine once you graduate), I'm going to be impatient and want the old style of problems. And I need more than just a mention that "trust me, you'll need this some day". Ideally, students will have lots of formative assessments that let them see how their abilities to do this kind of thing are improving over time.

D, E, A - Develop, Explore, Act an Understanding of the Problem

There are a lot of issues here--especially for me as a novice. Here's a few.

If you have a hip replacement with pain on one side, does the cane go on the side of the replacement or the opposite side (e.g. to balance the body). If you use crutches, would you stay off the leg on the side of the sore part of the hip or stay off the other leg? If I had pain in my hip I could easily experiment and see what helped best to ease my pain. But as a bioengineer, I eventually need to understand why things work as they do.

As I thought through the preceding questions it eventually became clear to me that I need to know much more about hips and how they support weight and how and why they can be painful. And I need to understand weight distribution in my feet as I walk, and how that affects hips. So at this point in my exam I might ask for a diagram on hips, etc. and receive several representations--some of which are and are not helpful. (I might be given this information by asking orally & receiving it via paper, or eventually via computer). If I don't ask for it, I might eventually be given it anyway and prompted about how to use it (this would be noted as part of the assessment of my performance). Eventually, I would be asked to calculate various things and say how they impact my solution to the overall problem above.

Look and Learn: I might be asked to say what I learned from this exercise that will help me improve next time.

Linking Assessments with the Concept Of Adaptive Expertise and New Views of Transfer

At a general level, the approach sketched above relates to adaptive expertise and new theories of transfer and their implications for assessment.

In my earlier thought piece on "adaptive expertise" that was sent around, I argued that most assessments are inadvertently built on a "sequestered problem solving:" (SPS) model of transfer rather than on a "preparation for future learning" (PFL) perspective (see Bransford & Schwartz, 1999).

Arguably, people who are adaptive experts are well prepared for future learning. They have developed the knowledge, skills and attitudes that allow them to frame problems and identifying what else they need to know in order to solve them. Most assessments, whether of memory or “transfer”, do not assess critical aspects of problem framing and searching for additional information. Instead, the assessments really look only at the end products of problem solving (e.g., the ability to compute given a very well defined task).

The latter is very important, of course. But we also need to assess the “front end” of problem solving. And we need to teach it and help students understand WHY it is useful to them in their lives. It would be great to do some mini-studies on this issue so we can show data to NSF by next year.

Appendix B
Design Methods for Evaluation of VaNTH Materials
David Cordray

Evaluation Design Options

The purpose of this section is to list the types of evaluation designs that might be used to assess the effectiveness of a module, mosaic, or course. These are enumerated from weakest to strongest designs for attributing observed effects to the module, mosaic or course. The key conditions necessary for each “design” are also noted. These designs can also be combined in various ways. For example, it is possible to incorporate a performance threshold within a randomized, two-group before-after design. As another example, multiple pre-tests can be introduced to obtain a better estimate of natural group trajectories within each group or condition.

One Group Designs (Innovation only)

1. ***Performance Threshold:*** This approach involves the specification of a numerical value that represents a level of performance that is *judged* or considered indicative of mastery or expertise. For example, a score of **90** percent on a test of biomedical engineering principles might be established as a threshold. Any student that meets or exceeds this level of performance is judged to be an expert. A model might be judged effective if 80 percent of the class met or exceeded this value. Both thresholds must be established *before* the module is presented and the post-module assessment is conducted. To avoid setting the criteria arbitrarily too high or too low, the criteria should be based on prior experience or group consensus. Over time the criteria can be altered (upward) to induce an expectation that there will be a continuous improvement as students and faculty get accustomed to this scheme.

This design is weak because the criteria can be arbitrary, grading can be biased in favor of students “passing” the criteria and there is *no basis* for determining how well the students would have done on the assessment without the module. The design can be improved if the same criteria have been used in the past (prior classes that were not exposed to the module, etc.) **and** prior assessments can be *re-graded* using a common scoring protocol. In this way, the prior performance, absent the module, can be used as a basis for comparison. Under this scenario, the design no longer involves one group. Cohort or group differences would need to be assessed and statistically controlled.

Please note that retaining final exam solutions over a period of time could help validate this method.

2. **Norm-based:** This approach is similar to the performance threshold design but has the advantage of using an agreed upon normed or standard for establishing expertise, understanding and so on. Such a norm might be the cognitive representation of major principles by adaptive experts in biomechanics. Students would be asked to conduct the assessment exercise as the expert and their responses would be graded relative to the responses of the expert. To judge if the module was effective in producing more “expert-like” representations, it would be necessary to establish a “closeness” or similarity criteria (e.g., the students reproduce 75% of the experts concept map) and a criteria for the group performance (e.g., 80 percent of the students met or exceed the similarity criteria).

Whereas the performance threshold design is weak because of the arbitrariness of the criteria for performance, this design relies on an external standard (performance of an expert, population-based performance of talents students, and so on). Weaknesses also include the arbitrary specification of the group performance and differences in expert performance (unless multiple experts are used and an average protocol is derived). The design is also weak because of the absence of basis for determining how well the students would have performed in the absence of the module (etc.).

3. **Before-After Change:** Rather than assessing student performance after the presentation of the module (etc.), this design requires the **addition** of a parallel assessment prior to the presentation of the module. By parallel we mean that the pre-test and the post-test must have the same properties (types of items, number of items, the same content is covered, and so on). If these conditions hold, the difference between the score on the post- and pre-test for each individual represents the maximum amount of change that *might* be due to the innovation. The extent to which the observed change is statistically meaningful is determined through the application of a statistical test (e.g., a correlated *t*-test).

Adding pre-module assessments improve the design but cannot account for confounding factors like normal learning, development or maturation, and the effects of practice in performing the assessment.

Two (or more)-Group Designs (Innovation and Control)

There are a host of research designs that involve comparisons between alternative conditions or groups. The estimate of effectiveness is represented as the **relative difference** between the two groups (on average) on the post-test assessment. If the post-test mean for participants in the innovative condition (the module) is 95.0 (with a standard deviation of 10.0) and for participants on the post-test is 85.0 (standard deviation =10.0), the relative effect is 10.0 units ($95-85=10$) or $95-85/10 = 1.0$ standard

units (or effect size). Whether the 10 unit difference is statistically meaningful is determined by conventional statistical procedures (e.g., *t*-test).

This section reviews some of the most common types of designs. Depending upon the target of the evaluation (i.e., a module, mosaic, course, curriculum) each design will be differentially applicable. It is unlikely, for example, that a “pull-out” design (see below) will be applicable in assessing HPL versus traditional courses. Rather, the pull-out design is best suited to smaller-scale innovations (e.g. modules).

Experimental Versus Quasi-experimental Two-Group Designs

In the following sections we present a variety of ways in which control groups can be established. Some of these are true experiments and some fall short of being true-experiments; we refer to the latter as quasi-experiments. Although the designs sometimes look alike, the main difference between these two classes of designs is the presence or absence of random allocation of units (e.g., students) to conditions; experiments employ randomization and quasi-experiments do not. This is not a trivial distinction. By using random allocation of units to conditions we can control the influence of subject attributes on the resulting outcomes. Without randomization we cannot be sure that individuals in each group are comparable (on average). For example, if we allow volunteers to enroll in an innovative module and give non-volunteers normal instruction (the control condition), any relative effects that we observe on the post-test are confounded by the influence of volunteerism. When the basis for self-selection into groups is known, we can measure individuals on these features and use statistical procedures to control for the effects of these factors. Sometimes these adjustments are successful; often they are not. The bottom line is that whenever groups are composed through non-random processes, our ability to attribute the observed relative effects exclusively to the innovation is reduced. Whenever possible, random allocation of students to conditions is the preferred tactic. Tables of random numbers appear in almost all statistics texts and can be easily used to assign students to groups.

To be effective in controlling for confounding factors randomization must be properly conducted and maintained throughout the experiment. If some students (non-randomly) fail to complete the post-test measure (also known as attrition from measurement), this introduces a bias. When attrition is substantial or differential across the groups, the groups become non-comparable and the initial randomization cannot be counted on to produce unbiased estimates of relative effects.

Timing of Measurement/Assessment

As noted above, performance can be assessed *after* the innovation has been presented (**post-test, only**) and it can be assessed **before and after** if a pretest is undertaken prior to exposure to the instructional conditions. With randomization the pretest is not necessary because random allocation of students to groups will equate groups (on average). But, there are several technical and practical reasons for

incorporating a **before-after** assessment sequence in the design. First, if randomization is used, pre-test assessment is desirable because it allows us to control for individual differences in student ability and aptitude prior to exposure to the conditions. This increases the sensitivity of the design (increases statistical power). Second, if attrition does occur, its influence can be examined by analyzing the pretest scores for individuals who remained versus those who have left the study. When randomization is not used to allocate students to conditions, the pretest assessment is required. Without a pretest, most non-random, post-test only designs are not interpretable.

Design Options

Most of the options for devising a control group can be undertaken using random or non-random allocation procedures. Rather than duplicating the list for random and then for non-random designs, we have discussed the distinction within each design option. In all cases, random allocation is the superior approach.

Historical Control Group: As the name suggests, historical controls involve groups that were composed in the past by non-random processes (e.g., students enrolled in the Spring 2001 section of BME 101). If adequate assessment were undertaken last Spring, the performance of students in the Spring course, presuming that the course entailed conventional instruction) could be used as a historical control group for an innovative BME 101 course that is planned for this coming Fall semester.

A weakness of the historical control group design is the possibility that composition of the cohorts can differ from one semester to another. Further, students in the later semester might be more experienced in the major and potentially possess more sophisticated thinking or problem solving skills. As such a battery or pretest measures are needed to assess the extent to which the groups are comparable. Sophisticated statistical techniques are required to adjust for cohort effects. And, it is unlikely that prior assessment processes will be adequate to assess the effects of the innovative (Spring) course. Without comparable assessment that is relevant to the anticipated effects of the innovative condition, it is not possible to test (in full) the effects of the innovation.

If administrative control over the assignment of students to semesters is possible, a stronger Fall versus Spring comparison could be developed. Because students do not always remain in assigned course (due to switching, dropping the course, and so on), these types of designs can degrade to quasi-experiments.

Delayed Control Group: We normally view the “baseline” comparison as a condition that occurs before trying the innovative strategy. If the investigator is willing to revert to the traditional mode of instruction in a subsequent time period, the “baseline” performance could be assessed after the delivery of innovative strategy. For example, the innovative material could be presented to students in the Fall semester, fully assessed with sensitive testing material, observations, and

so on. By switching back to the traditional mode of instruction in the Spring, the same assessment material could be used to examine student performance on measures relevant to the innovative strategy. This avoids the problem of having to plan all assessments prior to delivering the innovative strategy. As above, cohort effects need to be assessed and controlled, if possible. Maturation and normal growth would still be confounded.

Concurrent Control Group: Time-based confounding factors (e.g., maturation, experience, normal growth) can be controlled if groups or conditions created at the same point in time (i.e., they run concurrently). There are several options:

- i. **“Pull-out”:** In a class that enrolls a sufficient number of students, it is possible to create an intervention group (and in turn a control group) by dividing the class into 2 or more groups. Students could be exposed to innovative material in a lab or other setting, away from the traditional class. In other words, some students would be temporarily “pulled-out” of the traditional class. Both groups would be tested using the same assessment protocol. Using random selection would result in the most unequivocal results. If the total class size is small (e.g., 30 or fewer) the comparison might be underpowered (from a statistical point of view). The statistical power for any comparative design depends on the anticipated size of the effect of the innovation, sample size, and other technical factors.
- ii. **Other Courses:** When there are multiple sections of the same class in a given semester (**within the same institution**) composing a comparison group is one of the most desirable scenarios for developing comparison groups, especially if students have been randomly assigned to sections. If not, a pretest battery should be used to assess the comparability of the groups. It is conceivable that students with differing learning styles and motivations could choose one type of class or another. These factors would need to be considered as part of the statistical model used to assess relative effects. When the comparisons are developed **across institutions**, differences in selectivity of institutions also needs to be measured and incorporated into the statistical modeling process.
- iii. **Waiting List Control:** It is often possible to develop a large list of students who are willing to participate in a study. When the number of students exceeds the number of slots available for instruction (e.g., a lab that accommodates 20 students), the first set of slots can be filled by random selection from the list of volunteers. The remainder of the individuals would be put on a waiting list and offered the innovative lab after the first group (cohort) completed the lab. Members of both groups would be

pre and post-tested during the first administration of the innovative lab; members of the waiting list group would be tested after they received the innovation.

This design works best when students in both groups cannot interact and when the learning effects are expected to materialize quickly upon exposure to the innovative material. If the effect is delayed, additional post-test assessments are required.