

ASSESSING
THE
ELEPHANT

David Eubanks

Coker College

Copyright 2004-6 David Eubanks. All rights reserved.



CONTENTS

1. INTRODUCTION.	1
2. MOTIVATION.	3
3. THE EFFECTIVENESS LOOP.	5
4. ASSESSMENT.	11
5. COMPLEXITY.	18
6. IMPLEMENTATION.	24
7. RESULTS.	34
8. CONCLUSIONS.	43
APPENDIX	



1. INTRODUCTION

[T]he hardest thing about science is what it demands of us in terms of our ability to make the right choice in the face of incomplete information. – Lee Smolin, *Three Roads to Quantum Gravity*

Since Spring 2003 Coker College has been assessing student abilities in thinking and communications skills using course-based evaluations. This paper describes the philosophy, implementation, and results of this approach. It is my hope that it will inform institutional decision-makers about alternatives to standardized testing for judging the institution's liberal arts mission.

Although I have led this project, the faculty and administration at the College have been involved in every stage of development and made the actual assessment happen. They are wonderful colleagues and I am fortunate to be able to work with them.

The scope of this paper is fairly broad, including theory, application, and results. First we look at motivational factors in Section 2. Why are we in the business of assessment to begin with? Then, in Section 3, the abstract effectiveness loop, where assessment is used to make decisions, is described. Some potential problems are identified, and I define acts of defection that can compromise assessments. Section 4 addresses the assessment problem, particularly how language can influence our perception of outcomes. The difference between monologue and dialogue plays a large role, and some definitions are presented that will inform later discussions. In Section 5 we investigate the connection between assessment and complexity, including the apparent trade-off between reliability and complexity. The following two sections describe how the actual assessment program at Coker College was designed and implemented, and what the results have been. You can skip straight to Section 6 if you just want to know the design and results of the assessment process.

There are many examples throughout, which I hope will give flesh to the abstractions and help you find applications. Above all, I hope to provide some practical tools for thinking about the very complex business of assessing learning outcomes.

This material was first presented at the 2004 Assessment Institute at IUPUI. Late at night in the hotel room I was going through my presentation in my head and realized that one of my main points was the power of dialogue. It struck me as absurd to deliver a monologue about the virtues of dialogue, yet I only had 30 minutes to do the presentation—not long enough for real discussion. So the next morning my wife and I went shopping for laser pointers. During the session she helped me hand out 8 of them randomly to those attending my talk. Many of my slides had text on them such as "assessment of learning outcomes must be rigorous, reliable, and understandable." Then I asked my "pointers" to put their red dot on the adjective that they felt was most important. Using this

instant polling made the talk much more interactive without consuming lots of time. You may find ways to adapt this technique yourself. There are, of course, classroom technology solutions for this sort of this (personal response systems, they're called), but you can easily improvise one in this way.

The title of the paper comes from the fable of the blind men investigating an elephant, the object of a John Godfrey Saxe poem reproduced below.

Parable of the Blind Men and the Elephant
—John Godfrey Saxe

It was six men of Indostan
To learning much inclined,
Who went to see the Elephant
(Though all of them were blind),
That each by observation
Might satisfy his mind.

The *First* approached the Elephant,
And happening to fall
Against his broad and sturdy side,
At once began to bawl:
"God bless me! but the Elephant
Is very like a wall!"

The *Second*, feeling of the tusk,
Cried, "Ho! what have we here
So very round and smooth and sharp?
To me 'tis mighty clear
This wonder of an Elephant

Is very like a spear!"

The *Third* approached the animal,
And happening to take
The squirming trunk within his hands,
Thus boldly up and spake:
"I see," quoth he, "the Elephant
Is very like a snake!"

The *Fourth* reached out an eager hand,
And felt about the knee.

"What most this wondrous beast is like
Is mighty plain," quoth he;
"Tis clear enough the Elephant
Is very like a tree!"

The *Fifth*, who chanced to touch the ear,
Said: "E'en the blindest man
Can tell what this resembles most;
Deny the fact who can,
This marvel of an Elephant
Is very like a fan!"

The *Sixth* no sooner had begun
About the beast to grope,
Than, seizing on the swinging tail
That fell within his scope,
"I see," quoth he, "the Elephant
Is very like a rope!"

And so these men of Indostan
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right,
And all were in the wrong!

The Moral:

So oft in theologic wars,
The disputants, I ween,
Rail on utter ignorance
Of what each other mean,
And prate about an Elephant
Not one of them has seen!

The appendix at the end is another parable, this one written by David Kammler. It humorously illustrates what can go wrong in an effectiveness loop.

If you would like an electronic copy of this paper, you may download it from the web at <http://www.coker.edu/assessment>. I am interested in your feedback, which you can send to me at deubanks@coker.edu. I also blog about higher education at <http://highered.blogspot.com>.

Revision history:

10/28/2004 Original printed copy for the 2004 Assessment Institute at IUPUI

10/30/2004 I fixed some typographical errors and other minor cleanup. The only substantial change is the addition of a conclusions section at the end of Section 8. All pagination is the same.

11/23/2004 I added the text of the poem found above, and deleted the final section. A separate section on assessment philosophy is in the works, but hasn't arrived yet.

4/25/2006 Introduction edited, in preparation of a complete update with 2005-6 findings.

5/30/2006 Added 2005-6 findings and did quite a bit of editing to make the paper more concise, including eliminating the last section and replacing it with a short concluding section. Pagination has changed.

6/5/2006 I included some additional validity statistics based on NSSE 2005 scores. This can be found starting on the bottom of page 38. Somehow the pagination has produced two page twos, but it seems pretty harmless. This is the second page two.

6/9/2006 Added more NSSE information based on an analysis of variance comparison of means. I also corrected the legend of a graph comparing GPA to FACS scores and added a more detailed graph.



2. MOTIVATION

Q: What's the first question to ask?

A: "What's the first question to ask."

Q: What's the second question to ask?

A: "What's the second question to ask."

... no other types of questions are possible without a contradictory answer to the first question.

--The Decider's Paradox

The first question we really should ask is "why bother?" What difference does it make whether or not we assess students' critical thinking skills? I suspect you have your own answer to that question, but it's useful to start from first principles.

Accrediting bodies for postsecondary schools are explicit about the requirement for assessing educational goals. As a consequence of a research-based process, the institution is required to show mission success.

Example The North-Central Association requires a School Improvement Plan¹ that "Provides a specific assessment system designed to document increased student success on the goals identified."

Example The Southern Association of Colleges and Schools requires² that "The institution identifies expected outcomes for its educational programs and its administrative and educational support services; assesses whether it achieves these outcomes; and provides evidence of improvement based on analysis of those results."

We must assess to be accredited; without accreditation we don't get federal aid; without federal aid we must look for new jobs. Therefore we must assess.

The logic of the preceding paragraph is compelling but unsatisfying. Perhaps we can turn to the Council on Higher Education Accreditation (CHEA) for more motivation. They say³:

[I]t is imperative for accrediting organizations—as well as the institutions and programs they accredit—to avoid narrow definitions of student learning or excessively standardized measures of student achievement. Collegiate learning is complex, and the evidence used to investigate it must be similarly authentic and contextual. But to pass the test of public credibility—and thus remain faithful to accreditation's historic task of quality assurance — the evidence of student learning outcomes used in the accreditation process must be rigorous, reliable, and understandable.

There's some good advice here. Collegiate learning is indeed complex, and contextual authentic evidence is the most valuable kind. It's also a worthy goal that the evidence be understandable, and to the extent possible rigorous and

¹ www.ncacasi.org/standard/postsec/sip

² www.sacscoc.org/PDF/PrinciplesOfAccreditation.pdf, Section 3.3.1. I argue in my [blog](#) that this requirement is ambiguous and incomplete.

³ All CHEA quotes are from <http://www.chea.org/pdf/StmntStudentLearningOutcomes9-03.pdf>

reliable. However, the motivation given—that this is required to ‘pass the test of public credibility’—is probably optimistic since about half of the US population believes in ESP, presumably without rigorous and reliable tests⁴. Trying to convince the public of something sounds like a public relations job anyway.

Fortunately, there are more compelling reasons for assessing skills. An article from *The Economist*⁵:

Despite building and rebuilding elaborate growth models, [economists] have failed to prove that better education and training significantly raises a country’s long-term growth. Recently, though, a Canadian team made a breakthrough. It found that, if you measure actual skills rather than educational qualifications, human capital becomes a strong predictor of economic growth.

Some economists argue that ‘talent’ is over-rated, and it only takes a good attitude and patience to become skillful—emphasizing the role of education⁶. There are more specific indicators that skills are important, such as the College Board’s recent publication *Writing: A Ticket to Work... Or a Ticket Out*.

This seems like common sense—it makes a difference if an individual is able to perform complex tasks. We should therefore include measurement of the individual student’s skills in our assessment activities.

Following CHEA’s advice, we should look for a way to generate authentic evidence of complex skills and assess them in an understandable way. Before investigating that question in detail, we will take a quick look at the whole effectiveness process, of which assessment is only one component.

⁴ Source: Gallup Organization “American’s Belief in Psychic Paranormal Phenomena is up Over Last Decade,” (Princeton, NJ, 2001).

⁵ September 3—August 28, 2004 edition on page 70. I unfortunately didn’t record the title of the article.

⁶ *A Star is Made*, Stephen J. Dubner and Steven D. Levitt, New York Times Magazine may 7, 2006. What’s interesting to me about this idea is the role that motivation plays. We probably don’t do enough in higher education to try to change attitudes about learning. We’re too busy teaching.



3. THE EFFECTIVENESS LOOP

To err and err, but less and less.
-Karl Gauss

Many people have written about implementing institutional effectiveness, notably James Nichols, who has authored several books on the subject.

Institutional Effectiveness (IE) is supposed to be a continuous research-based process for making operational decisions. More specifically, practically every unit of the institution is required to set objectives, measure accomplishment (or lack thereof), and make changes based on the results. Further, the institution may be required to demonstrate that this process results in continual improvement relative to the objective.

One might imagine that effectiveness models are based on what one might loosely define as awareness. The schematic below is supposed to suggest a decision process of some hungry animal.

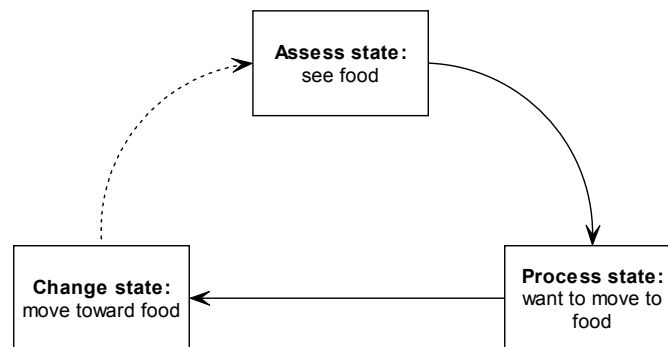


Figure 1. Hunger Decision Loop

At the top of the loop, our subject perceives some food. It constantly monitors its own state, and some states are more desirable than others. The state of seeing food but not eating it is not a good state when hungry. This information is processed by comparing the current state to possible actions that might change it (a formative assessment⁷). The conclusion is that moving toward the food is more desirable than simply sitting here looking at it. So the critter changes its state by beginning to move toward the food. This process repeats, showing continual improvement in progress toward its goal, until it's finally dining. At that point it has achieved its objective. Obviously this must be a grossly simplified version of what really happens, but it retains some elements of interest to us.

⁷ I will rely on the definitions of summative and formative assessment that can be found on pages 7-8 of *Assessment Essentials* by Palomba and Banta.

The effectiveness loop is a way to export to academic institutions a similar process. In order to complete the analogy to the critter above, we need to be able to assess, decide, and act.

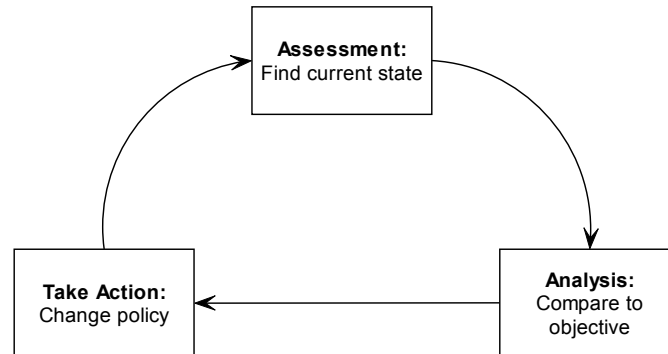


Figure 2. An objective's decision loop

The figure hides some of the details, but captures the essence of what's called the 'effectiveness loop.'

At this point, in the usual description of IE, it's supposed to be obvious to the reader that the recursive process shown in Figure 2, when properly employed, will lead to the accomplishment of objectives. We should be cautious about this conclusion, however. In physical systems all kinds of things can go wrong with feedback loops, like when a public address system squeals out of control due to microphone feedback.

Although there are many ways IE loops can fail, I want to focus on a particular problem having to do with assessment. If something is wrong with the assessment stage, it can happen that a loop *appears* to work, but actually accomplishes less than its objective. Worse, it's possible that the situation can actually worsen while the indicators get progressively more optimistic.

DEGENERATION

A stereotypical example of this effect is the argument that using standardized tests to measure learning objectives can actually reduce learning as the test scores increase, because the teachers focus on "teaching the test" rather than broader learning. I have no interest in wading into that debate, other than to discuss such situations generally.

I would like to call IE loops that fail while appearing to succeed *degenerate loops*. My tool for analyzing loops for potential degeneration comes from the economics subject called game theory⁸.

For our purposes, a "game" is an iterated process of interaction between an institution of higher learning and the stakeholders in the assessment

⁸ I find Richard Dawkins' *Selfish Gene* to be a good introduction to the basic concepts of game theory, although there are many books on the subject.

outcomes. Stakeholders may be students or faculty, or others involved. We will assume that the primary goal of the institution is accurate assessment. This is because we're after truth, not good news. In practice, individuals within the institution may well be happier with good results than they would with accurate results. This is especially true for stakeholders—those whose departmental budget, prestige, time, or money, may be affected by the process or its outcome.

In the process of assessment, the stakeholder may decide to “defect” in order to maximize one of the economic goals listed above.

Example A software developer found that there were too many bugs in its products, so it began a new system of rewards. Programmers would be paid a bonus for every software bug they identified and fixed. The number of bugs found skyrocketed. The champagne was quickly put back on ice, however, when the company realized that the new policy had motivated programmers to *create* more bugs so that they could “find” them.

An institution can seek to identify defection and account for it, or it can happily live with the flawed data it receives. In the latter case, it defects from its own purpose for having IE loops—it abandons the objective. It's convenient to summarize this in a matrix.

	Stakeholder Cooperates	Stakeholder Defects
Institution Cooperates	(1) Stakeholders make no attempt to manipulate data, and are fairly rewarded by the institution according to the specified metric.	(2) The stakeholder manipulates results to his/her advantage, and the institution rewards the manipulated assessment as if it were accurate.
Institution Defects	(3) Here the stakeholder doesn't manipulate the results, but the institution fails to deliver on the stakes, so the stakeholder is unfairly punished.	(4) The stakeholder manipulates data, but is not rewarded for it by the institution; the stakeholder is fairly punished for defection. The assessment is, however, invalid.

If the stakes are high to achieve convergence, this creates an incentive for stakeholders to produce the ‘right’ numbers. A degenerative convergence is the result of a measured outcome being steered toward target numbers by seeking to directly improve the assessment results, rather than improving the underlying phenomenon that the assessment measures. In this case the process appears to work when it doesn't. You might call this cheating, but labeling it as such may prevent us from understanding. The following examples illustrated some of the subtleties of these ideas.

Example If we use student evaluations as a major factor in assessing faculty, there will be pressure on faculty to get higher ratings, which they may do by a number of means, such as inflating grades. In extreme cases, the evaluations are faked by the faculty member and submitted as student responses, but much subtler forms of manipulation are common. One obvious thing that occurs to a faculty member is that assigning high final grades (which we know results in higher evaluations) is likely to be detected and perhaps investigated. So the instructor can give relatively high scores during the course, so that when a student fills out the course evaluation, she's satisfied with her current grade.

Then the professor gives a very hard final exam (after the course evaluation) and calculates the final grade by aggressively weighting it in the final average. The instructor gets higher evaluations but avoids being pegged as a grade inflator. The losers are the students, who were misled as to their performance during the semester, and anyone using the assessment results, who is being fooled.

Example If we base budgeting directly on assessment results, there will be a natural tendency to interpret results to justify new expenditures.

Example If grades are the basis for awarding scholarship dollars, grades will tend to artificially increase so that students can keep their scholarships.

Example US News rankings of colleges provide good press, so there's an incentive for schools to put their best face on reported results. This can take the form of selectively reporting SAT scores (do you report the students' highest, most recent, or average, scores?).

Example When I was in 5th grade, we had a box full of cards with math problems on them. It was self-scoring, and based on the score a student earned, easier or harder problems would be assigned for the next set. In retrospect I see that this was an attempt to guide me through a dependency tree of skills—in this case long division. Unfortunately, a student who was successful was 'punished' by having to do division problems with more digits than before, so after a while I 'defected' and intentionally put wrong answers to some questions so I could stick with the easier ones. The "institution" in this case was one Mrs. Ross, who *did not* defect, and I had to go back to doing them correctly. A year later cheap calculators became ubiquitous.

Example Students are summoned to a classroom to take a standardized test. It is explained to them that the test is an important measure of learning outcomes. Because their grade does not depend on the test, they have no reason to try to cheat on it. We might conclude that this is a low-stakes environment and defection is unlikely. However, students may well value their time more than they value some abstract test outcome they're unlikely to ever see. Some may decide to just fill in a few bubbles at random and leave. *The stakes only have to seem high to the stakeholder.*

The last example is a good argument for the advice from CHEA to use authentic data generated in some appropriate context. Defections at one level can affect decisions to defect at another level, as the following example shows.

Example A student fills in the 'bubbles' on a standardized answer sheet without looking at the test booklet. He's done in 2 minutes and wants to leave. Do you include that test with the rest, knowing that it's random and not representative of your students' abilities?

Example A student is tardy to class and misses the first 10 minutes of a standardized test. According to the protocol, the student should be excluded from the sample. But since this is a good student, there is an incentive for the instructor to allow him to take the test anyway. On the other hand, if the student had been an underperformer, the instructor could have proudly enforced the protocol and disallowed him from taking the test. Either way, the scores are improved.

While it is a natural reaction to decry manipulation of an assessment metric, it would be irresponsible to ignore these effects when calculating the 'actual' outcome of an assessment measure. Generally speaking, if a metric is

given an importance such that it affects job relationships, money, or other opportunities, it is given an economic value, which will be optimized by the stakeholders. A humorous look at this effect can be found in David Kammler's piece "The Well Intentioned Commissar," found in the appendix.

Just as credentials don't guarantee abilities, an IE loop doesn't necessarily produce improvements. Although the application of game theory to IE is a rich subject, I don't want to pursue it further here. The main lesson I wish to draw from this section is that we should avoid giving stakeholders incentives to defect. By this, I don't mean threats or punishments as a consequence, but by gathering evidence for performance in an authentic way and isolating assessment from policy-making as far as is possible. On the contrary, stakeholders should be involved with the assessment process and be interested in accurate results.

IMPLEMENTATION

Assessment results, meaningful or not, are presented to institutional policy-makers for action. There are at least two ways that theoretical effectiveness loops can fail at this point. Errant policies can derive from results that are invalid or good policies can be passed over because valid results are ignored. These correspond to Type I and Type II errors in hypothesis testing.

Example If SAT scores are actually a poor predictor of academic success, but the scores are nevertheless used in packaging institutional aid, this is bad policy based on a flawed assessment mechanism (type I error).

Example If research correctly shows that the cost of attendance is a good predictor of student attrition but this is ignored in setting financial aid policy, an opportunity has been missed (type II error).

A reasonable assumption about human behavior is that people do things that make sense to them. Applying this rule to institutional decision-makers prompts one to question the motivation and circumstances of actual policy creation. There are undoubtedly many factors that drive important decisions, including financial and political factors. Assessment results may actually be quite low in the list of motivating factors. This would caution any institutional researcher to ensure that policy-makers are very comfortable with the language used in assessment; that they can state questions and discuss results in routine vocabulary. If assessment results are cryptic or hard to relate to potential courses of action, they will likely be ignored. The more people who are conversant in the language of assessment, the more likely the results will be seen as meaningful. This is an argument for assessment being tied to an ongoing dialogue—a topic we will take up in the next section.

If consumers of assessment results are merely required to demonstrate use of the results, but are not motivated to defect by high stakes, this can create rather than eliminate options. A functioning institutional effectiveness system is then another organizational tool for use in administration, similar to a rolodex or

filing cabinet. A good leader can navigate the waters between Type I and Type II errors using good judgment. Assessment results will be ignored when they are not useful, and employed when they are, at least more often than random guessing. Just like in the stock market you need only be right 51% of the time to make progress.



4. ASSESSMENT

We say that a sentence is factually significant to any given person, if and only if, [she or] he knows how to verify the proposition which it purports to express—that is, if [she or] he knows what observations would lead [her or him], under certain conditions, to accept the proposition as being true, or reject it as being false. – A. J. Ayer, *Language, Truth, and Logic*

[T]he meaning of a word is its usage in the language. – L. Wittgenstein

How do we actually measure something like learning? The Council for Higher Education Accreditation advises⁹

What counts as evidence of success with respect to student learning outcomes is properly the province of each institution or program.

This presents us with the challenge to find an answer ourselves.

The two quotes above represent two almost incompatible epistemologies that correspond roughly to deterministic and intuitive approaches to knowledge. This is important because our design and use of assessment tools is affected by the beliefs we have regarding the attainability of knowledge. If we manage to deceive ourselves at the assessment stage, we may engage in meaningless activities like phrenology or astrology. Ayer would probably call such programs “Institutional Metaphysics.” I do not mean this to be derisive, since I’m quite sure that I regularly fool myself into thinking meaning exists where there is none. Even physical scientists can deceive themselves¹⁰.

For positivists like Ayer, if the meaning of ‘effective writing’ can’t be defined with some precision and an appropriate test, then the concept is literally meaningless. The problem is that we can’t *prove* one way or another whether effective writing is quantifiable in such terms or not. The creation of a technical description which then becomes the definition for the term is tautological: we must come up with the definition *a priori*. This is not merely theoretical. As Brian Huot notes in *(Re)-Articulating Writing Assessment*, creating a placement test for writing can *de facto* define what good writing is for the institution. A good grade on the placement test becomes synonymous with effective writing.

MONOLOGICAL AND DIALOGICAL DEFINITIONS

We will take an approach here that is more sympathetic to Wittgenstein’s statement. An artificial definition of writing effectiveness cannot survive in a setting where the idea of writing effectiveness can be debated, evidence examined, and a

⁹ All CHEA quotes are from <http://www.chea.org/pdf/StmntStudentLearningOutcomes9-03.pdf>

¹⁰ Smolin goes on to describe how Einstein fooled himself for two years in the development of general relativity with a wrong idea, and was rescued by his friends. I highly recommend reading the whole book.

consensus reached. This is the difference between a monologue and a dialogue¹¹. The following definitions will be very useful to us:

A dialogical definition is the usual meaning of a word found in language, which may change over time and be different for different people.

Example An English dictionary contains dialogical definitions. New words are born, old ones vanish or change meaning, and spellings change. The dictionary does not tell us what words mean so much as we tell it what words mean. It serves as an important repository for our knowledge of past usage, but it does not dictate how we use the language. There is no English language police. (France is a different matter.)

A monological definition is an enforced use of a concept that may or may not already have an established dialogical definition. It may explicitly define a word or phrase, or implicitly define categories.

Example I cannot get new glasses in my home state without undergoing an eye exam because of the laws governing eye exams. If I do get an eye exam, my vision is *defined to be* whatever the exam says, a monological definition. So despite the fact that I can see fine with my current prescription, a new one was forced on me, resulting in weeks of frustration. "To see fine" means different things to me than it does to the law.

A monological assessment is one in which the subject is subjected to a pre-determined regimen of tests that produce a discrete result. It is monological in the sense that communication of new information is unidirectional, from the subject to the administrator. It must also be used to categorize the subject in some way, enforcing a (possibly implicit) monological definition.

Example A basketball coach will quickly categorize players into those that can be centers and those that can't, based on how tall they are. The result is discrete (capable or not capable), and is enforced by the coach.

If an artificial standard, by way of a monological definition, is enforced in order to give a simple and measurable definition to an objective statement, it can lead to a bifurcation of meaning. A monological 'official' meaning, and the pre-existing dialogical common usage can exist simultaneously. For example, legal terms often have two meanings, and sometimes we use an adjective to specify the meaning, for example 'legally blind' and 'legally drunk' are the monological meanings (later we will use an asterisk to denote the monological meanings: blind* and drunk*). These kinds of meaning shifts permeate our everyday language.

Example The phrase 'to have a seat' is a common figure of speech to denote temporary ownership of a chair, as in "The bus was crowded, but I had a seat." If you substitute 'plane' for 'bus' the whole meaning changes. You can't get on a plane unless you 'have a seat' already, because of airline requirements. The airline defines what it means to have a seat: you have to pay first, usually a couple of weeks in advance.

¹¹ Although I know some of these ideas resonate with M. Bakhtin's literary theory, I am not very familiar with it. There is also a connection to De Saussure's distinction between *Langue* and *Parole*.

If we were to define a complex human trait like ‘intelligence’ or ‘creativity’ to be a number generated from a monological assessment, we neatly sidestep what is normally referred to by those words in everyday speech, and subordinate the broad and encompassing dialogical meanings. Such an external definition necessarily changes the meaning of the language in contexts where the new meaning is enforced. This is bound to create tension between the official and “folk” meanings. Rene Magritte’s painting “This is not a pipe” is a visual example of this tension.

Example You arrive at the airport gate too late—the plane is pulling away. You failed the performance test of getting to the gate on time. In your mind you are still a “passenger,” but to the airline, which is enforcing its rules, you are no longer a passenger. This creates a tension you’ll resolve by conforming to the airline’s definition (finding another plane and getting to the gate on time).

Example I got a score of 60% on my first Calculus test when I was a freshman at Southern Illinois University. The purpose of the test was to establish how much we had learned, and constituted a monological definition of skill attainment. At the end of the period where the professor handed the papers back, he said “What happened? You’re a better student than that.” Despite the fact that the test was his own devising, he trusted his own experience from classroom interactions rather than the test score, which categorized me as a dolt. Interestingly, he sought my opinion on the matter, too. In other words, his natural instinct was to engage in dialogue to resolve the tension between the score and his observation.

The last example puts a fine point on our dilemma as the creators of assessment instruments. We may create assessments that do not completely measure what we really want to measure. Our effectiveness plans may then lead us to the specified objective, but it may not be where we really want to go. It is easy to be fooled into thinking a metric is the “true” measure of something when it’s actually not.

Example In a publish-or-perish world, the common meaning of “scholarship” may be displaced by the number of publications a scholar has, which is easy to compute. Once this new definition of scholarship is in force, the stakeholders adjust by publishing more, and the proxy (number of publications) begins to drift farther from the “real” meaning of the word. Some institutions rate the quality of the publications in an attempt to restore some dialogue to the measurement.

Example An airplane’s artificial horizon, when working properly, gives the accurate orientation of the pitch and roll of the aircraft. In *Unfriendly Skies* by “Captain X”, the author describes two pilots flying in clouds who were amazed that both artificial horizons were malfunctioning. This was reconstructed from the black box recordings. Here the pilots took the dangerous step of trusting their inner ear assessments instead of their instruments.

Example For some reason peahens love big tails in peacocks. Over time, this fetish has led to the proliferation of genes that produce magnificent tails in the males of the species. Unfortunately, these tails aren’t good for anything except attracting mates. The rest of the time, they hinder the bird’s chances of survival.

The subject of sexual selection from evolutionary biology includes many fascinating examples like the last one, where a “local” definition of fitness (appealing to a mate) conflicts with a more general kind of fitness (day-to-day survival)¹².

These definitions will be useful to us in the development of later sections. They also give us convenient handles for different kinds of methods that we use in teaching.

Intuitively, a dialogue seems to give us more flexibility. It seems easier to miscommunicate with an email than with direct speech, for example. We investigate this idea next.

COMMUNICATING WITH MONOLOGUE AND DIALOGUE

Let’s explore the difference between a monologue and a dialogue when we want to communicate information. Imagine a trip to the doctor’s office where the doctor turns out to be out on vacation. “No problem,” says the receptionist, “Just fill out this form—it has every possible symptom listed on it. Our computer can read the results and tell you what’s wrong with you.” You look at the thick sheaf of paper the receptionist hands you. It contains page after page of yes/no questions like “Do you have trouble breathing when you lie down?”

How many questions would be necessary to diagnose all possible conditions? It would be a long list indeed, considering how many parts of the body there are and the variety of ways in which they can fail. Just covering allergies would take many pages.

Now consider a more normal visit to the doctor. You engage in a dialogue. If you complain about your knees popping, the doctor isn’t likely to ask you a lot of questions about your ears to diagnose the problem. There are many fewer questions involved because as each question and answer resolves a contingency, it eliminates many other potential questions.

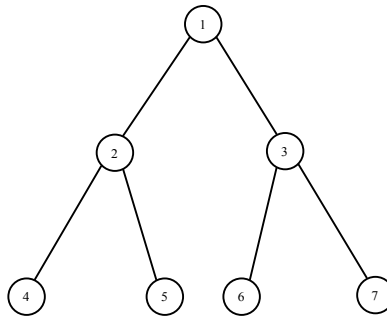
Many educational assessments have the common trait that they are a one-way communication. By this I mean that there is no recursive question and answer or other iterative process that constitutes a dialogue between the test giver and the test taker. So most tests are monological, with the respondent restricted to a finite vocabulary through his or her choice of responses. This limits the quantity of information communicated to a fixed number of bits¹³.

Example A 100-question instrument with four-response items communicates a maximum of 200 bits of information (2 bits per question). With a 1024-word vocabulary at your disposal, this equates to a very short paragraph (or a short German sentence) of 20 words.

¹² I highly recommend *The Selfish Gene* by Richard Dawkins for more on the subject.

¹³ Bit is short for binary digit. It is defined to be the information content of one yes/no item. It is the standard measure of information. Because there are two possible values for a bit, many information theoretic computations are most convenient in base two. A student in a literature class for the first time may have trouble providing more than one bit of analysis for a piece of literature: “I liked it” or “I didn’t like it.”

We can visualize the difference between the monological communication and the dialogical one with tree diagrams. The tree is composed of states, which represent knowledge of some event. The intermediate states represent partial knowledge, and the terminal states complete knowledge. This abstraction can be used as a metaphor for assessing the skill level of a student.



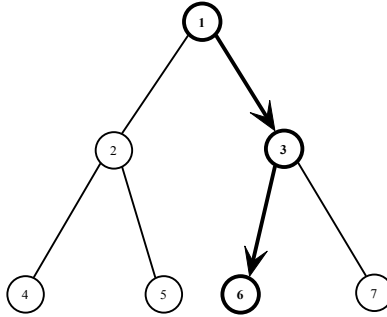
The tree in the diagram represents four distinguishable states, labeled 4, 5, 6, and 7. The other nodes are states of incomplete knowledge. That is, state 1 is the state of knowing nothing. State 2 contains knowledge of one bit of information (namely it rules out half of the final states: 6 and 7). With a monologue, we must ask a question for each of states 1, 2, and 3 because we cannot anticipate what the answers to those questions will be.

Example In case you need a concrete example, here is a possible list of questions corresponding to states 1, 2, and 3 along with the interpretation of the results made when the test is scored:

1. Compute 11×12 (if correct, ignore the answer to number 2, else ignore the answer to number 3)
2. Compute 3×4 (if correct, we assess state 5 to be the situation, else state 4 is)
3. Compute 145×256 (if correct, we assess state 7 to be the situation, else state 6 is)

At the conclusion of this three-question test, we can categorize a student's ability to perform basic multiplication. States 4, 5, 6, and 7 represent increasing ability levels. State 4 means the subject didn't demonstrate the ability to perform even the most basic multiplication problem. State 7 means the subject answered our most difficult question correctly.

For the same dependency tree with the same questions, a dialogical approach can skip the irrelevant questions, so that only two are needed to reach the maximum tree depth, as shown in the example diagramed below.



In the diagram, the answer to the first question ruled out states 4 and 5 as possible outcomes, so there is no reason to ask the question at state 2 which discriminates between those two.

Example With dialogue, we make decisions at each of states 1, 2, and 3:

1. Compute 11×12 (if correct, ask question 3, else ask question 2)
2. Compute 3×4 (if correct, go to state 5, else to state 4)
3. Compute 145×256 (if correct, go to state 7, else to state 3)

The rationale is that if a student can correctly perform an intermediate calculation, we assume she can also do an easier one, and challenge her with a harder one. This is the kind of thing routinely done in oral exams and independent study or coached courses.

It may not seem dramatic that the difference between the two approaches is one extra question for the monologue. However, if we generalize this to larger numbers of states the advantage of the dialogue is overwhelming. To traverse a context-dependent binary tree with n levels, a monological approach has a worst-case number of questions of $2^n - 1$ while the worst-case number of questions for a dialogical approach is n .

Example If the tree is 20 levels deep, as in the game of “20-questions” a dialogue can of course traverse it in 20 questions. To identify the same number of end states, a monologue would require $2^{20} - 1 = 1,048,575$ questions. I once bet a friend that I could identify any word in the Oxford English Dictionary in 20 questions. I would have worked, too, if he hadn’t chosen “Boticelli,” which isn’t in the dictionary.

Of course, not all questions are dependent on other questions. Sex and zip code may be independent in your data, for example. If *all* questions are independent, there’s no need to customize questions on the various branches of the tree—questions at the same level would all be the same. In this case, the tree analogy isn’t really needed, and it’s more useful to think of the questions as lying on different dimensions. In most cases there will be a combination of dependent and independent questions. The conclusion that dialogue has a powerful advantage over monologue is true in these general conditions¹⁴.

¹⁴ There are a lot of complexities I’ve overlooked for the sake of brevity. Like the fact that even a good student may blow an easy question, so some redundancy is desirable. In the end, however, the ability to omit questions you already know the answer to is still a powerful advantage.

If it's a human being directing the dialogue, this opens the process up to the criticism of being subjective—a different person might guide the dialogue down different paths if that possibility exists. That is, subjective measurements are criticized because they are hard to reproduce, and therefore probably have lower reliability than monological assessments, which are easier to make objective since information only flows in one direction. We will take up this topic again in the next section. In the meantime we should not forget that these particular subjective measurements are potentially *exponentially* more powerful at discernment than monological ones. We should especially consider dialogue when faced with our most difficult assessment challenge—that of assessing complex skills. This reinforces the CHEA's advice to rely on authentic evidence of performance.



5. COMPLEXITY

[I]f one has ten pounds of axioms and a twenty-pound theorem, then that theorem cannot be derived from those axioms. –G. Chaitin.

Some phenomena are harder to measure than others. Testing vocabulary in a German class is a good way to find out if students know the words, and it's likewise easy to find out if students can solve quadratic equations by giving them a list of problems. Determining whether or not a student exhibits analytical or creative thinking seems to be a more difficult task.

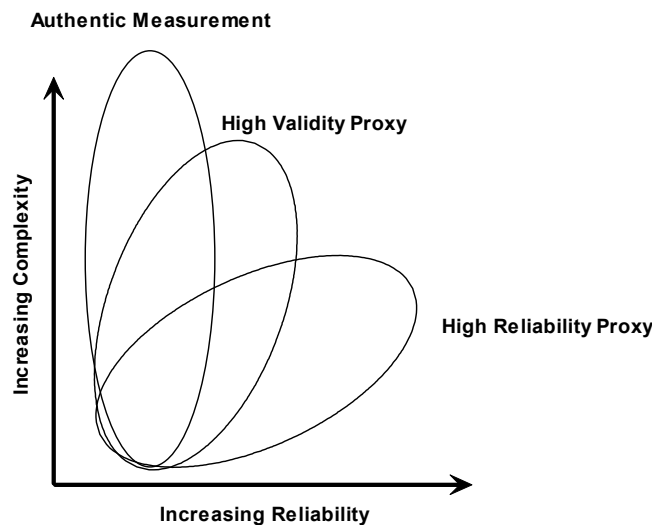
COMPLEXITY AND RELIABILITY I

We can define complexity explicitly and work from the definition, but that would take us too far afield. I will simply assume that some measurement tasks are more complex than others, and that the simplest ones are based on definitive rules.

Intuitively, simpler assessments can be done with more reliability. Aplomb and Banta describe this tradeoff between complexity and reliability in *Assessment Essentials* on page 89:

An [...] issue related to the reliability of performance-based assessment deals with the trade-off between reliability and validity. As the performance task increases in complexity and authenticity, which serves to increase validity, the lack of standardization serves to decrease reliability.

We can picture these relationships between skill complexity, reliability, validity, and authenticity as overlapping areas, shown below.



An authentic measurement is one that uses authentic evidence. A proxy is a substitute for authentic evidence. For low-complexity measurement tasks, we can find authentic and highly reliable tests. As the performance task becomes more complex, authentic measurements can't be reliably assessed because there is more disagreement among raters about the level of skill demonstrated¹⁵. The decision to give up authentic data in favor of a more standardized proxy or accept lower reliability instead, depends on what we want to achieve.

Example Writing samples from class work can be considered authentic in that context. But it's hard to compare one student's paper on mitosis with another student's paper on Milton. To increase reliability, a timed essay can be used. This is an artificial proxy for the authentic data, but it has the advantage that it can be designed to make it easier to score. Reliability can be increased further by administering a true/false test on grammar and other writing skills. This is a low-validity/high-reliability proxy. As we'll see, we've changed what we're measuring from writing to writing*, which in this case gives us reliable, but partial, information.

Example Suppose we want to fill an open staff position. We can imagine sending a written questionnaire to candidates, which would have a certain amount of reliability (the candidate is likely to answer in similar ways in repeated trials if we choose to conduct them), which we could increase by using a 1-5 Likert scale for each question. Probably no one would hire a candidate based solely on such a response, however, because it has fairly low validity—the candidate may fool us into believing things that are not true. An interview has higher validity because it involves a dialogue and human interactions—sure to be an important component of any position. The interview is not very reliable, however, because a different composition of interviewers is likely to gather different data and give different subjective ratings. A six-month 'trial period' of actually *doing* the job is likely to be a highly valid indicator of ultimate job performance because it's authentic data. We can't repeat the first six months as a confirming experiment, though, so there's no way to check actual reliability¹⁶.

Example A quantitative example of this effect at Coker College is the use of indicators to predict first-year GPAs of entering students. High school transcripts are a reasonably good proxy because earning grades in high school is similar to earning grades in college. On the other hand, SAT scores, no matter what their reliability may be, have lower predictive power for our students' first year GPAs. In other words, the scores are a low validity proxy.

This presents us with a kind of Uncertainty Principle. For complex performance skills, we can have authentic assessments or we can have reliable ones, but not both together. This makes sense, because reliability itself comes from a predictable cause-and-effect relationship that is not a trait of complex systems like higher-order mental activity. As a consequence, we should be

¹⁵ This is what the diagram indicates, but that doesn't make it true. A speculative argument for why this should be so is found in the last section.

¹⁶ You may argue that the second six months of work performance is closely correlated to the first six months, and we can define reliability that way. That isn't a fair comparison though. While we can imagine convening a new interview committee to check results, it's hard to imagine creating a whole set of supervisors for the second six months of work. It's a debatable point, though.

prepared to give up a certain amount of reliability when we assess complex phenomena.

On the other hand, we value many learning outcomes that are not complex at all. These include learning the rules of algebra and the rules of grammar, spelling, and punctuation. Much depends on students reliably solving rules-based problems, and we can easily test these with monological instruments like multiple-choice exams if we choose to.

ASSESSING THE ELEPHANT

The “Turing Test” is a famous example of measuring complex behavior related to the mission of the liberal arts. We would like to validate that our students are *thinking*, not simply applying some memorized rules and facts in a rote fashion. Our conception of the general education skills is as a set of related abilities that enable higher-order problem solving that cannot be imitated by, say, a machine. That is exactly the test that Alan Turing¹⁷ designed to judge whether or not a computing machine could think. His test is to have a human correspond remotely with other individuals through something like an Internet chat service. One of the correspondents is actually a computer program being judged for intelligence. The computer takes input from the human and generates responses. The test is this:

If a computer, on the basis of its written replies to questions, could not be distinguished from a human respondent, then ‘fair play’ would oblige one to say that it must be ‘thinking’¹⁸.

Turing chose a *dialogue* as our most complex measuring instrument—no number of exams, timed essays, or multiple-choice tests could ever persuade us that a machine was intelligent. Only an interview that demonstrates something more complex than algorithmic thinking has the power to (possibly) convince us¹⁹.

Example The following thought experiment contrasts with the statement above. Suppose a designer promises that his new computer is intelligent. As proof, he takes a standard IQ test and a number two pencil and feeds it into the machine. After a few moments, the paper returns with every item correctly answered. Would we consider this as evidence of machine intelligence?

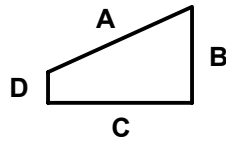
The very usefulness of freedom of thought from constraints makes it hard to precisely measure such freedom. “Analytical Thinking” is viewed differently in Art from the way it is in Mathematics. If we were to narrow the definition enough to get total agreement, we would have lost the ability to discern the very

¹⁷ Turing wrote the paper that founded modern computer science.

¹⁸ The quote is from *Alan Turing: The Enigma*, by Andrew Hodges

¹⁹ I do not wish to express an opinion on machine intelligence, merely to draw attention to the importance of the distinction between simple algorithmic and complex behavior. The Turing Test is conducted every year, and you can read about it at <http://cogsci.ucsd.edu/~asaygin/tt/test.html>. There is debate about whether or not the Test is sufficient to prove intelligence.

phenomenon we were trying to measure. The diagram below helps illustrate the idea. Four observers A, B, C, and D contemplate a tall solid object.



If we ask the observers how big the object is we get four different answers. Only by engaging in a dialogue can they reach an understanding of the various dimensions. The perspectives of the raters represent contexts that cannot be averaged out.

Example A red octagon containing the word STOP is effective writing at a street intersection but probably not on a resume.

The observers could simplify the situation by just agreeing that D is the ‘correct’ answer, but this replaces understanding of the actual complexity with a (more reliable) simpler proxy. This is fine if we remember to distinguish later between the proxy and the “real” assessment. The proxy represents only a part of the picture.

Confusing the whole and the parts is called a mereological fallacy²⁰. In the story of the blind men exploring the elephant, they each feel a different part of the animal and through dialogue conclude what it is. If the one who is feeling the leg were to bully the others into accepting his “tree” solution, we could hardly call the result of the singular measurement a description of an elephant. Instead we would have to qualify the language by putting an asterisk and write elephant*. Given several specimens, leg measurements would reveal differing characteristics, and could be used to classify them. Still, the description of the leg is not a description of the elephant. We are instead describing elephants*.

In order to avoid using an asterisk to indicate partial information, we must assess *every aspect that can be perceived*. There is no point in talking about aspects that *cannot* be perceived.

Example If we assess grammar and spelling, we are not assessing writing, we are assessing writing*. If we insist on a writing assessment that requires high reliability, we are restricted to those aspects of writing that can be reliably measured. We are again measuring writing*.

Example What exactly is it about reading the ‘Cliff’s Notes’ version of a novel instead of the actual item that seems like cheating? Intuitively, a student who does this misses out on the complexities of the original, although he or she may well absorb some of the important details, particularly those likely to be useful on tests and papers. In the notation we’ve introduced, the student is reading *Odysseus** or *War and Peace**.

²⁰ In *Philosophical Foundations of Neuroscience*, pg 74, Bennett and Hacker give examples of this kind of error, and I am using their notation. A similar approach would probably turn up examples in assessment literature where complex skills are being measured.

If we *enforce* a view that skill* = skill, we create a monological definition, which creates a tension between it and the usual (dialogical) meaning, as we saw in the previous section.

Example Years ago as department chair, I tried to implement a point system for faculty annual evaluation, where research, teaching, and service were assigned levels as objectively as I could manage. That wasn't a very popular idea with some because it seemed to oversimplify the matter. In retrospect I agree.

Assessing everything sounds impossible, and it is if you think of the whole as a sum of parts: writing = writing* + writing* + ... This reductionist approach may work once we really understand how the brain works, but we're nowhere near that stage currently. It's time to define a complement to monological assessment.

A dialogical assessment occurs when the evaluator uses interaction and observation to reach conclusions about a subject, and reports those in natural language.

Note that any assessment that is used as a categorization tool enforces a monological definition. In this case, it is the judgment of the evaluator that implies the definition.

Example An interview with a prospective employee that results in a decision to offer or not offer a job is a dialogical assessment. The employer implicitly or explicitly defines what it means to be hireable in that position.

Example Assessing writing by assigning papers, collecting them, and grading them is monological. Allowing students to submit revisions based on your comments is dialogical.

The idea is to assess complex skills in the same way that we teach students: through language with a constructive dialogue.

At the end of the previous section, I promised to address the subjectivity of the measurements. Subjectivity sometimes gets a bad name because the results of subjective measurement seem not to be reliable (in the sense that multiple observers may reach different conclusions). We will partially address that problem by requiring our observers to be experts in what they are measuring. A biologist is an appropriate person to judge analytical thinking in a biology class, for example. Does this assessment match an analytical thinking assessment in an art class by an art professor? In Section 7 we show reliability results.

Conceptually we are much better off by staying close to common usage for terms like "analytical thinking" than we are by creating monological definitions. Philosophers and even scientists²¹ often try to have it both ways, and "define" common terms more precisely so that they can be used to reach concrete conclusions. The problem is that the conclusions only apply to the new definitions, not to the common usage of the words.

²¹ See *The Philosophical Foundations of Neuroscience*, by Bennett and Hacker for a good discussion of this.

A successful theory can create new and useful definitions for old words. A good example is Newton's use of the word "force," which in his time had not the technical meaning it does today in physics. But it's clear that if we say "force of will" or "force for good" we are talking about the common usage, and if we say "force equals mass time acceleration" we are talking about the technical usage.

Without developing a full explanatory theory of learning that is on par with Newton's revolution in physics, I don't see how it's possible to define, let alone test for, the cause and effect processes that allow demonstration of complex human skills. Therefore, rather than choosing reductionism and standardization to assign performance levels to complex behavior, we should consider using complex instruments, such as the holistic subjective judgments of intelligent observers.

SUMMARY

We can create something we call an "intelligence test," but this isn't how we really define intelligence. We hope that it's a good proxy for what we mean by intelligence, but if we have to decide between our own judgment and the results of the test, a reasonable person will trust his own subjective assessment.

Language is fuzzy, and this is a utility. We can see a kind of dog we've never seen before and still happily categorize it as a canine. Fuzziness is also a problem because it makes it difficult to reach concrete conclusions. We therefore invent special languages for tasks where precision is required. If we do that, however, we must create the *whole* language, like logic, mathematics, or physics, if we want to avoid collisions with common meanings of words. It is a sophist's trick to use technical language in common speech to add the appearance of logical inevitability. Our approach at Coker College in implementing complex skills assessment is to avoid reductionism and standardization when possible. We turn to that next.



6. IMPLEMENTATION

The system is never rong.

-System documentation 1.0²²

To summarize the main points so far:

1. We are motivated to teach skills to students because those skills are valuable to the students after graduation;
2. To that end we employ a recursive “observe and act” process of improving their education;
3. The process is not guaranteed to succeed, but we can avoid obvious pitfalls by:
 - a. Engineering assessment so as not present motivations for defection among stakeholders;
 - b. Building assessment processes appropriate to the task. For complex skills, a dialogue is the most discriminating method, although it may be the least reliable;
 - c. Avoiding artificial definitions of outcomes;

The elements of Coker College’s Faculty Assessment of Core Skills (FACS) are built on these ideas.

THE FACULTY ASSESSMENT OF CORE SKILLS

Our implementation consists of the following elements:

1. The skills to be assessed are analytical thinking, creative thinking, effective speaking, and effective writing;
2. Our objective for each skill is similar to the one for writing:

When Coker College alumni are supervised in an academic or work environment by educated people who have themselves demonstrated writing skills in obtaining an advanced degree, these supervisors will judge that the writing skills of Coker College alumni are appropriate to a college graduate.

²² This statement is ironic on two levels. The word ‘rong’ is coined in *Canman* by Stanislaw Zza. It means “not even wrong,” to use Wolfgang Pauli’s famous phrase. Wrong means we didn’t get the correct answer. Rong means we weren’t using the right method to get it.

3. For each course taught, the instructor will:
 - a. Decide which of the skills will be assessed in the course;
 - b. Write rubrics for levels of achievement for each skill assessed. These are published in the syllabus. The levels are:
 - i. Graduate. This student has achieved the level of skill we expect our graduates to demonstrate;
 - ii. Junior-Senior. This student has achieved a level of skill appropriate to an upperclassman.
 - iii. Freshman-Sophomore. This student has achieved a level of skill that we expect of underclassmen.
 - c. Assign a rating to each student for each of these skills at the end of the term. Ratings do not affect grades in any way. Data are entered on a web-based form. Two additional rating options are present on the form:
 - i. Blank. No rating assigned for this student.
 - ii. Remedial. This student performed below the Freshman-Sophomore level. (A rubric for “remedial” may or may not be present in the syllabus. It’s not required.)

4. Scores are aggregated and reported:
 - a. As a consolidated report to the college community;
 - b. By major to the department chair;
 - c. By student to advisors;

5. The institutional effectiveness committee discusses the results, with input from other campus committees, and decides what action is required, if any.

Not all of these components were implemented at the same time, and as of this writing, the system for feeding back information to advisors is still incomplete.

Some comments on each of the preceding items:

1. Creative and analytical thinking, and effective speaking and writing are presumed to be complex skills.
2. This is a dialogical goal. That is, we are not seeking to put a new meaning to the words “effective writing.” Also, we are interested in our students being *perceived* as good thinkers. We will test that with the proxy of our own perceptions.
3. The rubrics on course syllabi are general guides only, and are specific to the course. Assessments are not restricted to a single in-class assignment. Rather, instructors are encouraged to base the final rating on the whole class experience. Ultimately they have to decide for themselves what constitutes assessment. Ratings are *not* for improvement, but level

of demonstrated skill, according to the rubrics. Ratings are independent from grades.

Rubrics from a Mathematics course are shown below as an example.

Core Skills Assessment	
<p>The goal of Coker College is to graduate students with the ability to think analytically and creatively, and to write and speak effectively. Students will be during the semester as to their demonstrated ability in the following areas. The results are not used to determine grades, but rather are used to help the College improve programs and better advise students.</p>	
Analytical Thinking	
<p>We will think of analytical thinking in this course as the skill of mastering rules and applying them. There are many definitions, theorems, and procedures involved in learning linear algebra. These include the notions of dimensionality and span of a vector space, and the processes of reducing matrices to a particular form, or performing algebra with these mathematical objects. Some of these are easy to master, and some require more sophistication.</p>	
Graduate Level:	<i>Can check correctness of proofs for results that are unfamiliar, use novel notation, or advanced concepts. Understands eigenvectors/values, vector spaces, and inner product spaces and can solve straightforward problems on paper and using Maple..</i>
Junior-Senior Level:	<i>Can correctly apply matrix multiplication and do matrix algebra, understands the notational differences between matrix algebra and scalar algebra, understands use of transpose, inverse, and commutivity properties. Understands and can check correctness of basic proofs.</i>
Fresh/Soph Level:	<i>Master the mechanical processes of row reduction, solving linear systems, using a pencil and on Maple. Know basic properties of vectors and matrix algebra.</i>
Creative Thinking	
<p>We will think of creative thinking as the active use of imagination in solving linear algebra problems. This skill is exhibited when you solve a problem that has no pre-specified 'step-by-step' solution. It usually involves trial and error, and it's essential that you have mastered the analytical thinking skills necessary to know whether your work is technically correct or not. That is, creativity may include defining new objects and new rules, but cannot include breaking the rules unless it's to create a new 'superset' of rules.</p>	
Graduate Level:	<i>Shows facility to produce proofs for results that are unfamiliar, and use mathematical induction properly. Can apply theoretical understanding to applications. Can use Maple control structures such as do loops to create working procedures.</i>
Junior-Senior Level:	<i>Can apply intermediate linear algebra concepts to real-world situations and solve the problems correctly. This includes applications of matrix multiplication. Can prove theorems that</i>

Fresh/Soph Level:	<p><i>require Fresh/Soph analytical skills. Can write Maple procedures and debug them.</i></p> <p><i>Can apply basic linear algebra concepts to real-world situations and solve the problems correctly. Can use the help system in Maple to discover and use new capabilities of the software or debug problems.</i></p>
-------------------	--

The web-based form used to gather assessment data is displayed below.

[Read the instructions!](#) Click the submit button at the bottom of the page when you have finished the survey.

MUS 230-11
Which core skills have you measured in this course?

Analytical Thinking
 Creative Thinking
 Effective Writing
 Effective Speaking

*Please mark the skills measured in this course that were demonstrated by each student.
 If you didn't measure a skill, just leave that column blank.*

Student	Measured Achievement Level			
	Analytical	Creative	Writing	Speaking
Allan, Michael (0136753)	<input type="text"/>	Remedial	<input type="text"/>	<input type="text"/>
Basket, Clifford (0162487)	<input type="text"/>	Fresh-Soph	Fresh-Soph	<input type="text"/>
Charles, Frances (0021985)	<input type="text"/>	Jr-Sr	Graduate	<input type="text"/>

Notes:

See the course syllabus for rubrics. One student did not complete the paper assignment.

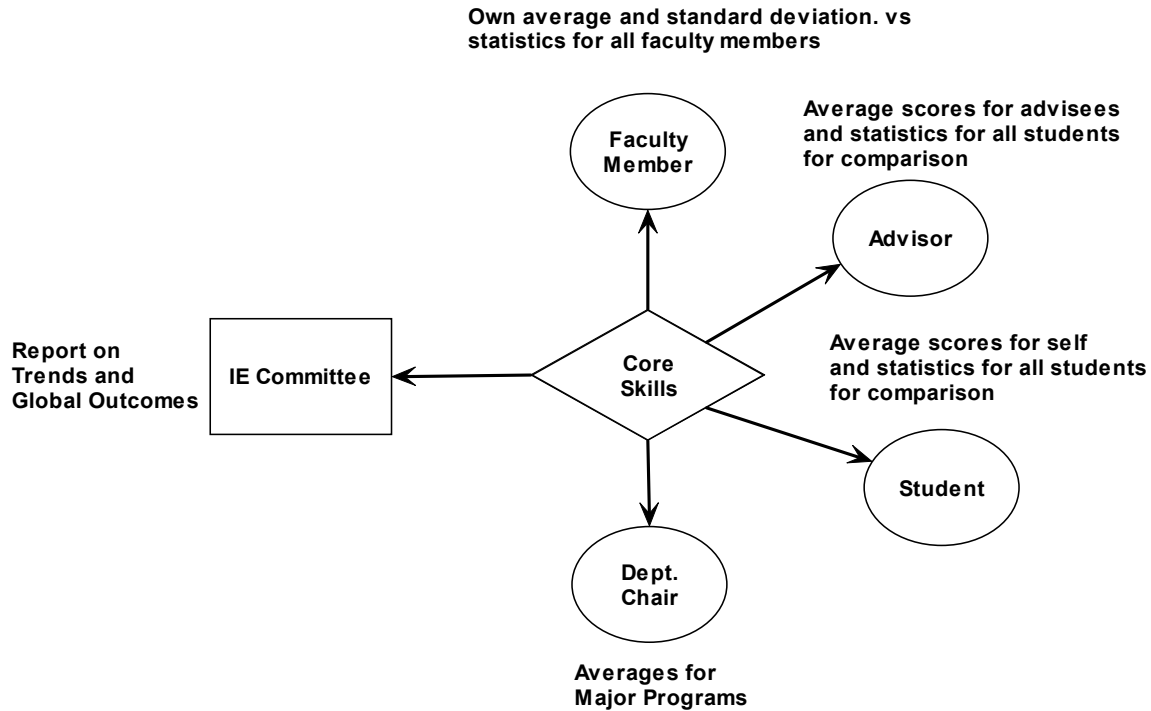
Click here when you are finished!

- The scores are averaged by assigning 0, 1, 2, 3 to the successive skill ratings, with remedial being the lowest. Technically this is non-parametric data and averaging is problematic²³, but so far that's what we've done.

²³ Thanks to Russ Hamby for pointing this out to me.

USING THE DATA

The purpose behind this design is to create a college-wide discussion of the skills in the context of class work. This introduces language about performance that is not related to grades. The aggregation of scores allows different constituencies to see data in different contexts. The schematic below shows who gets what information.



Institutional Research prepares a report, which is then published. The IE committee fosters discussion about the meaning of the results. Other committees also use the report for decision and action²⁴.

Advisors may get a report showing average ratings for the advisee, like this one.

AvgOfA	AvgOfC	AvgOfW	AvgOfS
1.29	1.25	1.14	1.50

It would be accompanied by overall averages for students in the same seniority level.

²⁴ We have a Liberal Arts Studies Program committee in charge of general education decisions.

Major programs (here called the 'Generic' major) receive reports that compare their students to those in other programs.

FACS for Generic Majors

Class	N	Analytical	Creative	Writing	Speaking
2001	38	1.67	1.70	1.84	1.96
2002	167	1.69	1.71	1.63	1.88
2003	88	2.00	1.88	1.90	2.05
2004	96	1.30	1.29	1.31	1.32

FACS external ratings for Generic Majors

Class	N	Analytical	Creative	Writing	Speaking
2001	22	1.65	2.00	1.94	2.00
2002	91	1.52	1.49	1.54	1.81
2003	30	1.83	1.75	1.71	2.00
2004	65	1.29	1.11	1.25	1.29

FACS for all other majors combined

Start	N	Analytical	Creative	Writing	Speaking
2001	239	1.42	1.55	1.31	1.46
2002	1216	1.48	1.48	1.43	1.54
2003	1691	1.31	1.34	1.28	1.36
2004	911	1.33	1.26	1.23	1.34

This allows them to compare ratings within the program to ratings from instructors in other disciplines.

SETTING OBJECTIVES

In the effectiveness loop, assessments are tied to goals. We'll compare two alternatives for setting goals based on a FACS-like assessment.

Tire Gauge Model

Goal: Ensure that students learn to think analytically.

Assessment: The FACS will be conducted as described.

Analysis: The institutional research office will compile average results.

Objective: The average FACS score for analytical thinking for seniors will be one level higher on average than when the same students were freshmen.

This objective looks good because it conforms to the usual standard for constructing such statements. The objective is arbitrarily, however. What exactly does it mean that seniors score a point better than freshmen? Aside from this, the objective has two more serious flaws:

- It creates an incentive for stakeholders to defect if are consequences to failure. If the raters know that seniors are “supposed to” score higher than freshman, they will comply, perhaps unconsciously.
- It does not allow debate about the meaning of the assessments. Once a number is enshrined as the annual assessment outcome, it takes on more importance than it really should have.

As we have seen, it is difficult to create a meaningful summative assessment of a complex skill. The objective gives a false sense of meaning in the same way it would if I tell you I’m 69.50% happy.

By its nature a summative assessment is bound to contain a very small amount of information, which is unlikely to adequately reflect the subtleties of the situation. Sometimes that’s okay—it’s fine for a tire gauge to average out molecular interactions to give you a pressure reading. You probably aren’t interested in the probability distribution of momentum for those particles. On the other hand, increasing tire pressure is a simple procedure: put more air in the tire. It’s not at all obvious how one “puts more air” into student minds so they can think better.

A way to address these deficiencies is simple, but not usual.

Telescope Model

Goal: Ensure that students learn to think analytically.

Assessment: The FACS will be conducted as described.

Analysis: The institutional research (IR) office will conduct research on the data to look for evidence for and against improvement in student skills. The report will be published internally and discussed in a faculty meeting. Following this, IR will produce an addendum that includes faculty commentary and seeks to answer questions raised in the meeting with elaboration, correction, or further analysis.

Objective: The objective will be considered a success if the final report compels most readers to believe that the goal is being achieved.

The modifications replace summative numbers with dialogue about the meaning of the results. This dialogue has the potential to identify complex patterns in the data, generate questions for future research projects, and suggest modifications to the design of the experiment. It’s impossible to say beforehand what will come of it. This is much like building a telescope: you have no idea what you’re going to find when you look up.

The tire gauge vs. telescope difference boils down to the difference between summative and formative assessment²⁵. Although we can create a

²⁵ I find these categories extremely useful. A lot of fog cleared for me when I read Palomba and Banta’s description in *Assessment Essentials* (pg 7-8), where they credit Erwin, Ewell, and Terenzini with the idea.

summative assessment for a complex skill, we should be cautious about its meaning. At the least, these should be complimented with formative processes (this is standard assessment advice). The tire gauge model looks like a scientific approach, but it's not. The important part of science is not the numbers and equations, it's the dialogue about those numbers and equations. There are always more special cases to be accounted for and explained. And there are mistakes in philosophy and methods that must also be detected through informed debate.

Imagine presenting evidence of achieving learning outcomes to a room of chemists, language professors, philosophy professors, etc., and asking the question "does this evidence convince you?" That takes courage because the answer may not be positive. But this "telescope model" engages the community in debate and gives them a voice and a stake not only in the outcomes *but in the methods* of assessment²⁶.

The formative "telescope model" has a weakness too. It assumes that the people engaging in dialogue care about the outcomes, ask intelligent questions, and are intellectually honest. Colleges and universities are a good place to find such individuals. In the end we depend on Socrates and the virtue of knowledge: if we can discern good from its opposite, we shall choose good.

DEVELOPMENT OF THE FACS

As the end of the Spring 2003 semester approached, IE meetings and meetings with academic departments had produced some parameters for a general education skills survey. We wanted to put a pilot implementation in place to see what problems we might have. So we built a web-based reporting system that allowed each faculty member to report a pass/fail rating for each student they taught and for each of the core skills. The standard for a pass was this: "Has this student demonstrated the level of skill we expect of our graduates?" The technical aspects of the survey worked reasonably well, although some of the data mysteriously vanished. A research report analyzing the results showed that perceptions of student achievement increased with the number of credit hours they had earned, which was encouraging. The instructors had inflated the scores to an unhealthy level, however.

I presented the results to our Council of Chairs. We had a good discussion of the results, and they made excellent suggestions for improving the survey. Based on their comments, we went to a four-level scale with a rating of "remedial" at the bottom, and "graduate" at the top, as was described earlier.

There were some technical glitches with the new forms, but the responses were very good and the data much improved over the first trial. I tried to look for

It's amazing how powerful a simple idea can be in dividing a chaotic jumble of concepts neatly into two orderly halves.

²⁶ In my experience, the number of new research questions generated in such debate far exceeds the capacity of the Institutional Research Office to answer. If you attempt this, be prepared to weed the list in some politically sensitive way. There is always a need for leadership.

longitudinal improvements in students' ratings, but the two different rating systems made this difficult.

At the end of the Spring 2004 semester, we had a full year's worth of data on the new scale. Moreover, we were making some progress on the conceptual front. The IE committee, with the Provost's help, engineered new requirements in course syllabi so that they would have to include rubrics for the various skill levels being assessed. This was fully enforced in the Fall of 2004, with all but one faculty member complying. During the first week of the Fall 2004 semester, the faculty participated in several days of workshops on understanding the FACS, writing rubrics for the core skills, and creating learning objectives. Despite an aggressive schedule, I was impressed with the positive attitudes I saw. A snap survey showed that faculty members appreciated the information about the FACS. Part of that session included a mini-lecture on testing reliability, and I distributed to each a sheet with the individual's own average rating compared to the global average.

It has become apparent that the FACS is more than an assessment tool—it also is a pedagogical device. Faculty have shared with me experiences in talking about the core skills rubrics with students. One said that her students' writing assignments have improved just because they now have a definition of remedial work. When I asked for feedback, one faculty member emailed me:

The introduction of the core skills language with regard to writing has helped me consciously create in-class activities that sample students' writing. When I review the samples outside of class, I can see what the majority of students do or do not understand. I can then make additional worksheets or activities to clarify specific points or to correct their weaknesses.

Another wrote:

Students seem less confused by the term "analytic"—at least it seems familiar to them now—and they no longer seem to call it "picking things apart." But I am not sure how much they have internalized at this point.

I do review with each class what the expectations are for each skill, and this is in addition to reviewing the rest of my syllabus (including evaluation criteria). So we spend more class time on this topic than I used to. . . .

In my own case, I began labeling problems in my linear algebra course as creative or analytical. The students adopted this language themselves, and can now usually tell the difference in the problem types themselves. I find it useful in assigning homework and creating exams to try to balance the analytical and creative skill problems to match their burgeoning abilities. On their exams I give them two scores: the total and the percentage of "creative" points they earned. The latter is for information only, but they understand from our discussions that it's an important part of their development as a mathematics major to be able to think creatively. My students now understand that the reason they find proofs initially difficult is because they are making a transition from a "follow-the-rules" analytical process for solving problems to a creative process that draws of all of their knowledge.

The FACS has had an impact on the culture of teaching and learning at the College, particularly in affecting our vocabulary. How far this goes and how long it lasts we will be able to tell in time.



7. RESULTS

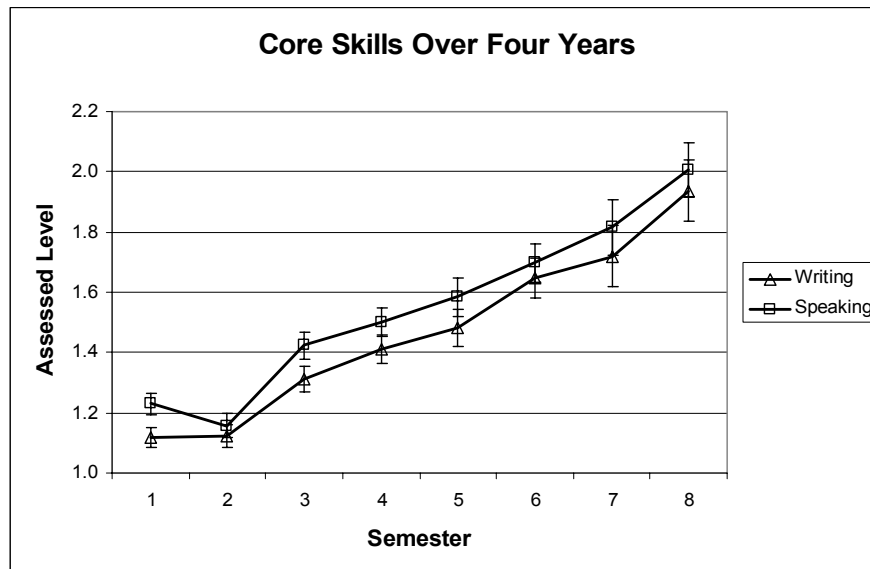
The only true wisdom is in knowing you know nothing.
-Socrates

In this section we look at the results of the Faculty Assessment of Core Skills (FACS) taken from the 2003-4 through 2005-6 academic years.

The 2003-4 academic year was the first full year of results from the improved Faculty Assessment of Core Skills (FACS). There were 4,501 responses, 2,733 from the Fall and the rest from the Spring. Each response is a rating of a student by a course instructor in one or more of the core skills. There were 1,023 students rated, or about 91% of the students taught. Counting the multiplicities of each skill for each student, there were a total of 19,733 assessments, or about 19 assessments for each student who was rated.

By the end of the 2005-6 academic year we had a total of 11,770 responses rating 1743 students, or 85% of the traditional student population during the 2003-2006 period. There are difficulties in getting adjuncts to comply with the assessment, which has lowered its usefulness in the evening programs.

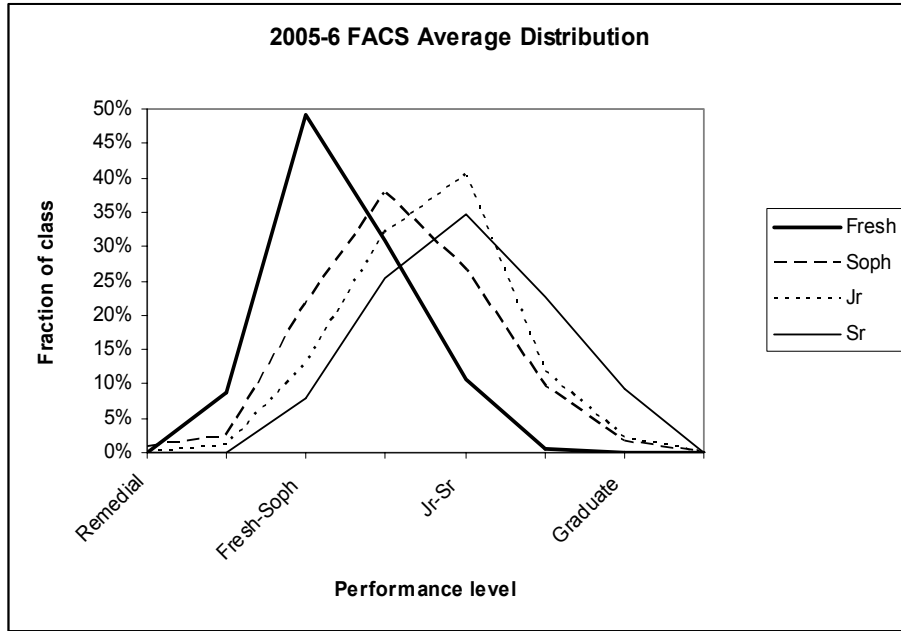
In presenting the results, I will use averages of ratings, although it may not be reasonable to assume that the difference between the levels is uniform²⁷.



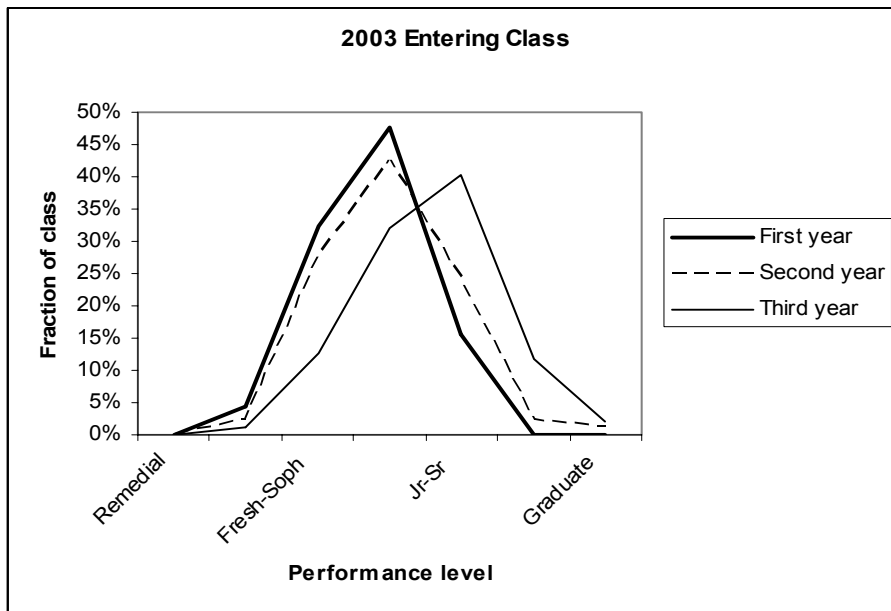
The graph above compares two of the four skills aggregated for classes beginning in Fall 2002 through Fall 2005. The other two fall between these and were omitted for readability. These longitudinal results track students in their initial semester through their eight, only including students who did not leave

²⁷ In other words, it's probably better to use medians rather than means. For this report, however, I don't want to give up the convenience of averaging results. Our reliability statistics are done using non-parametric measures.

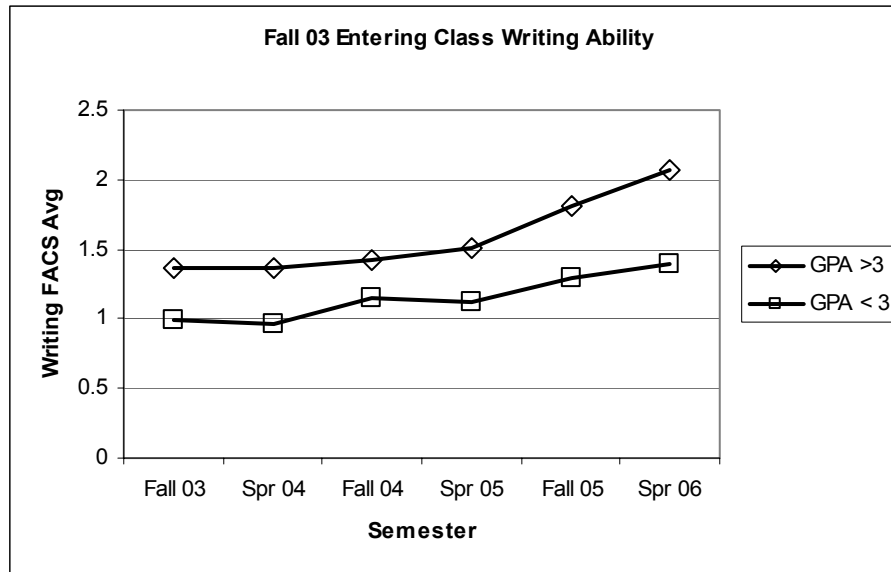
college during that period (otherwise the results might improve simply due to survivorship). The trend is upward, as one would hope. If we simply averaged the scores for the four skills, we would expect the distribution of scores to be roughly bell-shaped. The graphs below show such distributions for each class (freshman to senior) using 2005-6 score averages. They are based on scores from 459 students.



The means of the distributions are in the order we would expect. A longitudinal distribution for the 2003 entering class shows evolution over time.



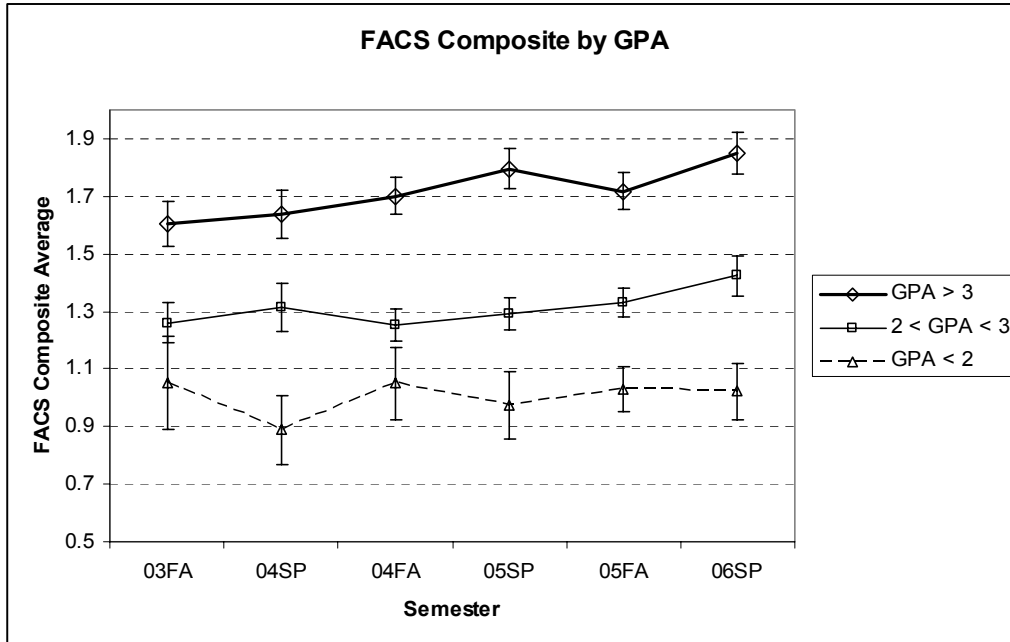
If we classify students by their overall grade point average (3.0 or greater, or less than 3.0), we can graph the average writing abilities of these groups.



This graph shows longitudinal data for those members of the Fall 2003 entering class who were still enrolled in Spring 2006. We can see that although both categories of students showed improvement in writing ability, after three years the least-performing group was only at about the level that the best-performing group started. Additionally, the gap between the two groups of students widens as if “the rich get richer.” This raises interesting questions about the ability of students to learn, and what we can do about it.

A more detailed examination of the effect of GPA classification on measured abilities compares composite FACS scores of students who were enrolled in Spring 2006 within bands of overall grade point averages. This is not truly longitudinal since not all students attended all semesters. Still, it clearly shows that the students who are in the lowest GPA class (<2.0) seem not to improve at all in demonstrated ability.

The error bars on the graphs are set at two standard errors.

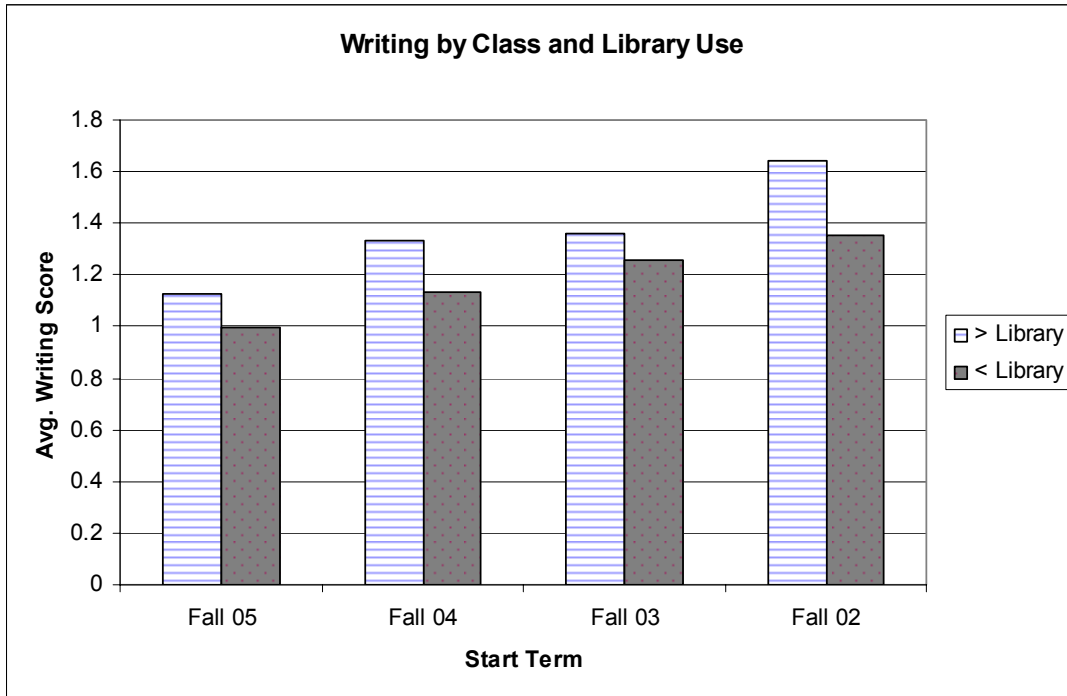


The clear relationship between FACS and GPA is a demonstration of external validity since the measured ability levels act the way we would expect them to. We turn now to other measures of this sort.

EXTERNAL VALIDITY

We can investigate validity further by comparing FACS results to other variables. We find, for example, that:

- Correlation between FACS analytical thinking and overall GPA is 0.64.
- Students who access library databases score significantly higher in effective writing than those who do not.
- Similarly, students who check out 10 or more books from the library (denoted > Library) score better in writing than those who do not (< Library).



The graph compares students with ten or more circulation events logged in the library database to those with fewer than ten. The data are as of Spring 2006, and show four entering classes (first through fourth year students).

In Fall 2005 we launched an electronic portfolio system for the purpose of capturing authentic student work. Near the end of the semester a group of faculty and invited guests assessed sampled student portfolios. Each was rated using a rubric in five categories. These are shown below with correlations to grades and FACS writing effectiveness scores.

Correlation Table N = 152

	<i>Purpose</i>	<i>Audience</i>	<i>Content</i>	<i>Arrangement</i>	<i>Style</i>	<i>GPA</i>	<i>FACS</i>
Purpose	1						
Audience	0.79	1					
Content	0.74	0.70	1				
Arrangement	0.78	0.76	0.77	1			
Style	0.66	0.67	0.70	0.71	1		
GPA	-0.16	-0.21	-0.19	-0.24	-0.29	1	
FACS	-0.27	-0.31	-0.29	-0.26	-0.37	0.65	1

The highest correlation with FACS was in the Style category, with a coefficient of -.37. The scale for the portfolio review was reversed from that of FACS, so the negative numbers are appropriate.

One also has to wonder if course instructors assign the higher scores to more senior students simply out of expectation that they will perform better. Do

seniors get higher scores simply because there are more opportunities to see students perform at higher levels in more advanced classes? In discussion with the Council of Chairs before the survey was administered, this question was asked.

It could also be that instructors are simply biased toward higher scores in more advanced courses. If so, does this account for the entire difference by semesters of credit we see in the graphs? A correlation, for Fall 2003 entering students shows that both time and course level are important. Here, time is measured by the year the class is taken for the student: first, second, third, or fourth. Course level is taken from the designation of the course: 100, 200, 300, or 400-level.

Correlations

		Year	Level	Rating
Year of Class	Pearson Correlation	1	.442(**)	.319(**)
	Sig. (2-tailed)	.	.000	.000
	N	642	642	642
Class Level	Pearson Correlation	.442(**)	1	.241(**)
	Sig. (2-tailed)	.000	.	.000
	N	642	642	642
Average Rating	Pearson Correlation	.319(**)	.241(**)	1
	Sig. (2-tailed)	.000	.000	.
	N	642	642	642

** Correlation is significant at the 0.01 level (2-tailed).

Using a partial correlation we can factor out the influence due to class level.

Partial Correlation of Year and Rating, Controlling for Class Level

Control Variables			Year	Rating
Class Level	Year of Class	Correlation	1.000	.243
		Significance (2-tailed)	.	.000
		df	0	639
Average Rating	Average Rating	Correlation	.243	1.000
		Significance (2-tailed)	.000	.
		df	639	0

The effect of time is still significant. For the 2002 entering class, it was also significant with a coefficient of .196.

Another set of data for comparison is the 2005 NSSE results. The sample size is somewhat limited, and after scrubbing the data I had a sample size of only 97 students in both Day and Evening programs. An average FACS score was calculated for each and compared to certain items from the NSSE that seemed to

be directly relevant to general education. In this study, the three significant correlations were:

Item 1a. *Asked questions in class or contributed to class discussions.*
Correlated with FACS at 0.21 with significance of .044.

Item 1f. *Came to class without completing readings or assignments.*
Correlated with FACS at -.28 with significance of .007. So students who reported that they were unprepared were somewhat likely to score lower on the FACS.

Item 11a. *Acquiring a broad general education.*
Correlated with FACS at .25 with significance of .023.

For this data set the starting year of the student correlated with FACS at -0.57 with $p < .001$. The only NSSE item of those examined that correlated significantly with start year was:

Item 1p. *Discussed ideas from your readings or classes with faculty members outside of class.*
Correlated with start year at -0.21 with significance of .035.

Using a different statistical method, I classified students by FACS averages as below or above the median average composite score. Then I compared average NSSE scores for each item for significant difference using analysis of variance. The items significant at the .02 level are:

Item 1f. *Came to class without completing readings or assignments.*
Higher FACS scorers reported doing this less often.

Item 6b. *Exercised or participated in physical fitness activities*
Higher FACS scorers reported doing this less often.

Item 9d. *Participating in co-curricular activities*
Higher FACS scorers reported doing this less often.

Item 9e. *Relaxing and socializing*
Higher FACS scorers reported doing this less often.

Item 13 *How would you evaluate your entire educational experience at this institution?*
Higher FACS scorers rated the experience higher.

RELIABILITY

One measure of the meaningfulness of the core skills assessments is our ability to reproduce results. That is, if there's something there to be measured, we should be able to measure it twice and get the same result. Of course this ideal experiment is easier to execute when measuring tire pressure than when measuring the creative ability of a student. Nevertheless, we must ask the question in order to establish the boundaries of what we know.

In order to get an idea of reliability, we could ask "what is the probability that two randomly selected instructors will agree in their assessments of a student they both rated?" This is relatively easy to compute, and we do not have to have the same number of samples for each student, as we would with ANOVA.

The equation for the probability of a match is given by:

$$p = \frac{1}{n} \sum_{j=1}^n \binom{m_j}{2} \sum_{k=1}^{m_j} \binom{s_{jk}}{2}$$

Where n is the number of students, m is the number of samples for a given student, and s is the number of scores of 0,...,3 respectively. In cases where $s < 2$, the factor is defined to be zero. Using a computer program we found the probability of two measurements agreeing. The results are shown below for each of the skills.

Analytical	50%	with 820 students sampled
Creative	53%	with 691 students sampled
Writing	50%	with 799 students sampled
Speaking	54%	with 646 students sampled

For comparison, the likelihood of a match of two student ratings taken at random²⁸ is 35.5%.

The probabilities in the table are reduced because we're asking for an exact match, and being close doesn't count for anything. A problem with this is that not all raters use the same scale. That is, they might *rank* 10 students the same way, but give them different ratings in absolute terms.

A more standard measure of reliability is the Interclass Correlation Coefficient (ICC). In order to calculate it, we had to pare the data down quite a bit to fit the protocol (four measurements per student in analytical thinking) to 258 students.

The ICC for this data is .426, which means that 42.6% of the variance in scores is presumably due to differences in the subjects, rather than differences in raters. This indicates that the assessment is a useful measurement. For

²⁸ I computed this by squaring the frequencies of each rating (0,1,2,3) and summing. You might assume that the result should be 25%, but that would only be true if the distribution were uniform. It's interesting that this number is the Euclidian norm of the distribution, if you think of it as a vector. It would be a good student research project to sort out why.

comparison, the model that we use to predict first year GPA at Coker explains only about 15% of the variance using high school GPA and SAT as predictors.

Generally speaking, the reliability of the FACS is much lower than what you'd demand of a standardized test. But I have argued that in judging complex skills we must be prepared to sacrifice some reliability in favor of validity.



8. CONCLUSIONS

I realized I was dyslexic when I went to a toga party dressed as a goat.
-Marcus Brigstocke

The FACS results show that perceptions of student abilities increase logically as we would hope. There are good indicators of validity, and reliability is acceptable.

Analysis of the results shows that the data seem sensible and in line with what we would expect. The resolution seems to be good enough to draw conclusions about disaggregated groups. Generally speaking it looks as though the FACS is achieving its purpose of tracking the core skill abilities of students.

The FACS seems to have shown its worth as a complement or replacement to externally defined standardized tests. We have not systematically assessed the attitudes of faculty, but as the administrator of the program I can say that they see it as an important task to be done at the end of a semester, not a burden or busy work. The mission of the College is also discussed much more (at least the part dealing with liberal arts) at all levels of administration.

Our approach will probably not work for everybody. Coker's emphasis on the individual student makes it natural for us to focus more on validity than reliability. In the end, we would like to be able to say of a graduate "she can think analytically, and can explain complex ideas well" instead of "she can think* at 83% and speak* at 91%". The size of an elephant is more than a number on a tape measure.



APPENDIX

The following is included with the permission of the author. From the perspective of the reader it's a humorous parable about a defective assessment loop. I'm sure it's not quite so funny to be on the inside of a story like this.

THE WELL INTENTIONED COMMISSAR

(A parable for academic workers and those who direct their activities)

David W. Kammler, Professor
Mathematics Department
Southern Illinois University Carbondale

In the early days of Communist Party rule in Poland, the State Central Committee ordered the Commissar for Building Materials to improve the productivity of the eight factories where iron was turned into nails. The Commissar was a well-trained bureaucrat who decided to implement quantitative management techniques (which he only partially understood).

I must carefully define a measure of performance that will enable me to identify and reward increases in “productivity,” he thought to himself. After some reflection he drafted a memo informing the eight Directors that during the coming year they would be evaluated by using the performance metric

$$P = \frac{\text{Number of nails produced during the current fiscal year}}{\text{Number of nails produced during the previous fiscal year.}}$$

The Commissar was filled with self-satisfaction as he anticipated the presentation he would give to the State Central Committee, precisely quantifying the annual productivity increases.

The ambitious (but unscrupulous) Director of the Wroclaw nail factory, Stanislaw Nowak, read the Commissar’s memo and immediately called his counterpart at Poznan. Dear comrade, he said, we are all being pressured to increase productivity and I know that you have been having difficulties with your outdated press for making boxing nails (the 2½" long 8d nails used to attach boards to studs and rafters when building a house). I have a surplus boxing nail press here at Wroclaw that I am willing to swap for your little used wire brad machine since mine is almost beyond repair. And so he traded

Wroclaw's only boxing nail press (the adjective "surplus" being a socialist white lie) for a second machine for producing wire brads (tiny ½" nails occasionally used for small trim work). He made a similar offer to the Director of the nail factory at Lublin, obtaining a third brad press in exchange for his machine for making framing nails (3½" long 16d nails used to fasten studs, joists, rafters, etc. when framing a house).

Throughout the year the Wroclaw factory concentrated on the production of wire brads. Workers who had previously been assigned to the boxing nail and framing nail presses now produced tiny wire brads. Workers who had previously produced roofing nails, finishing nails, etc. during the day were reassigned to the night shift where they made still more wire brads. At the end of the year, the Commissar carefully calculated the figure of merit for each factory. The nail factories at Poznan, Lublin, Gdansk, Szczecin, Katowice each had a noticeable increase in productivity with scores of 1.12, 1.11, 1.08, 1.05, 1.07. But by concentrating on the production of wire brads Stanislaw Nowak had achieved a productivity metric of 8.15, in spite of the fact that he had only processed 20% of his iron allotment for the year! The delighted Commissar rewarded the successful Director by assigning him a villa, and he rewarded each of Nowak's workers with a kilogram of bacon and a large basket of rutabagas.

As one might expect, the storage warehouse at Wroclaw was filled with keg upon keg of unneeded wire brads. Moreover, since the iron that should have been used to make boxing nails and framing nails was also sitting in the Wroclaw warehouse, Polish carpenters began to experience times when the supply depots ran short of the nails they needed most for house construction. The Commissar took note of the situation and reasoned as follows. I will ask the State Central Committee to "solve" the national nail shortage by increasing the iron allotment for each of my factories. (Commissars are always looking for an excuse to ask for additional resources!) Of course, the real reason for this year's slight shortage of boxing nails and framing nails is that in spite of its outstanding productivity the Wroclaw factory was unable to process all of its annual

allotment of iron. I can easily remedy the situation by changing my productivity metric to encourage the conversion of iron into nails. With supreme confidence in his analysis, he informed the Directors of his eight nail factories that for the coming year the productivity metric would be

$$P = \frac{\text{Kilograms of nails produced during the current fiscal year}}{\text{Kilograms of nails produced during the previous fiscal year.}}$$

The Directors of the nail factories at Gdansk and Szczecin invited Nowak to conduct seminars on “The Wroclaw wire brad strategy for productivity enhancement”. They were envious of the stunning success of their colleague at Wroclaw and eager to implement the methods that led to his recognition and reward. They were absolutely delighted when he “reluctantly” agreed to exchange two of his wire brad presses for their seldom used machines for making log spikes (8" long 80d nails used for constructing wood bridges, log cabins, etc.). The Director at Katowice made a similar deal, exchanging his log spike press for the Wroclaw tack machine. After all, he reasoned, tacks are only slightly larger than wire brads so the same management principle must certainly apply.

Throughout the year workers at the Wroclaw plant fed iron into the four log spike presses, keeping them running day and night. Time and again additional iron allotments were obtained from the State Central Committee to keep the factory humming. When the Commissar computed the annual productivity metrics, he was dumbfounded. The nail factories at Gdansk, Szczecin, Katowice that adopted the “tiny is better” strategy had the disappointing production metrics of .22, .21, and .26, respectively. The factories at Czestochowa and Krakow, with Directors who routinely ignored any and all management directives, had lackluster metrics of .98 and 1.01. But the factory at Wroclaw had a production metric of 21.30! This time Nowak was given a Polonez (Polish luxury car)

with a driver, and each of his workers received a kilogram of bacon and two large baskets of rutabagas.

The Wroclaw warehouse was now filled with keg upon keg of unneeded wire brads and keg upon keg of unwanted log spikes. Since only four of the eight factories were still producing boxing nails and framing nails, the shortage was now acute, with construction delays of 1-2 months being common during the critical summer building season. The State Central Committee wrote a formal memo to the Commissar, ordering him to investigate the situation and take immediate steps to correct the problem.

It's merely a matter of choosing the right metric, the Commissar said to himself. Since he had learned a bit of statistics during his school days he was able to identify precisely what had gone wrong during the previous two years. A productive factory must make nails of various sizes to serve the needs of our building trades, he reasoned, so we want the standard deviation of the nail size to be large, not zero as has been the case at Wroclaw during the past two years. A productive factory must also turn a lot of iron into nails. And in view of the current crisis it seems best to focus on annual production rather than some measure of year to year improvement. With this in mind he created the productivity metric

$$P = \left(\begin{array}{l} \text{Kilograms of nails} \\ \text{produced during the} \\ \text{current fiscal year} \end{array} \right) \times \left(\begin{array}{l} \text{Standard deviation of the weights} \\ \text{of all the nails produced during} \\ \text{the current fiscal year} \end{array} \right).$$

When this management directive was received at Poznan, Lublin, Gdansk, Szczecin, and Katowice, the discouraged (demonstrably unproductive!) Directors had no idea what it meant and no absolutely incentive to find out, so they decided to follow the example of their peers from Czestochowa and Krakow and simply keep on doing exactly what they had done the year before. They could not comprehend why the production of minuscule wire brads or massive log spikes was regarded as meritorious at a time when the Polish building trades were starved for common boxing nails and framing nails.

Meanwhile, the canny Director of the Wroclaw plant once again ordered his workers to maximize their production of log spikes throughout most of the year. Just before the year ended, he asked them produce an identical number of tiny wire brads. (This made the standard deviation of the nail weight approximately half that of a massive log spike.) And for the third year in a row, the Wroclaw factory was judged to be the most productive in all of Poland, with a performance metric 36.2 times larger than that of the next best factory. The Commissar publicly congratulated Comrade Nowak and his workers at the Wroclaw nail plant for a job well done, and bemoaned the fact that in these hard economic times he was unable to offer them even token rewards for their outstanding achievement.

Unfortunately, the national shortage of boxing nails and framing nails was now so severe that thousands of carpenters were unemployed and the whole economy was depressed. The members of the State Central Committee were greatly displeased with the Commissar. After all, they had clearly identified a problem and told him to solve it, but in spite of his best efforts things had gone from bad to worse. He was given one last year to remedy the situation.

Our factories must be made to produce nails that our carpenters can actually use, he reasoned. We cannot encourage them to make nails that sit rusting in their warehouses. With a sudden flash of inspiration he created a corresponding measure of performance

P = Kilograms of nails actually sold by the factory during the current fiscal year.

He winced as he realized that this metric reeked with the foul smell of capitalism, but he was desperate and could think of nothing else to try.

The Directors at Poznan, Lublin, Gdansk, Szczecin, Katowice, Czestochowa, and Krakow were shocked to receive a management directive that violated their most deeply

held Marxist convictions, so once again they ignored the memo and continued with the same production schedule they had used for the past two years. In contrast, the devious Nowak immediately recognized that the Commissar could not very well chastise him for adopting a capitalist response to an intrinsically capitalist directive. For the third year in a row, he ordered his workers to produce as many log spikes as possible and then left Wroclaw for a six-week holiday at an exclusive Black Sea resort. At the end of the year, he took these log spikes together with all of the log spikes and wire brads from his warehouse (i.e., virtually every nail that his factory had produced during the previous four years!) to a local firm that made tracks for the Polish railroad. He sold the whole lot at half the going rate for scrap iron. And this year, Nowak's production metric was 9.2 times larger than that of his closest competitor!

The State Central Committee summarily dismissed the Commissar for Building Trades and initiated a search for his replacement. Two weeks later, the Committee met to consider resumes of the applicants. Comrades, said the Chairman as he studied what appeared to be a most promising vita, I believe that I have identified a leader who can set us on a Five Year Plan that will rejuvenate our failing economy. During each of the past four years Stanislav Nowak, the Director of our nail factory at Wroclaw, has led his workers to be the most productive in our nation. I have never seen such consistently high measures of productivity. These numbers do not lie. Even in this severe depression, his workers are 9 times more productive than those at any other nail factory. And so Stanislav Nowak was named Commissar of Building Trades for all of Poland, a most remarkable honor for a man who managed a factory that had produced absolutely nothing but high performance metrics during the previous four years.

MORAL: When we lose sight of our purpose, we inevitably become unproductive, no matter what numbers we generate to prove otherwise.

Permission to copy: You may share copies of this story with anyone who might enjoy thinking about its implications in academia. Such samisdat publication (or to use the equivalent Polish phrase podziemna publikacja) is in keeping with both its setting and with its message.