Research and development of enhanced assessment tools for chemistry education

Thomas Holme
Iowa State University

ACS DivCHED Examinations Institute

PERC 2011, Omaha, NE

A fundamental challenge

- Teaching is, at once, inherently personal and inescapably corporate.
- At present, the corporate interests in student learning are largely articulated in terms of assessment.

Exams Institute?

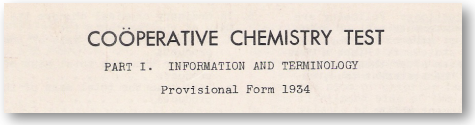
- How is it that chemistry has an Exams Institute?

A Short History (part 1)

- 1921-formation of Division of Chemical Education
- 1930 – formation of Committee on Examinations and tests
  - Committee subsidized by:
    - General Education Board of the Cooperative Test Service
    - Carnegie Foundation for the Advancement of Teaching
    - Dr. Ben Wood

A Short History – part 2

- 1934
  A group of five Committee members released the first general chemistry test in three forms


COÖPERATIVE CHEMISTRY TEST
PART I. INFORMATION AND TERMINOLOGY
Provisional Form 1934

A Short History (part 3)

- 1946 - Ted Ashford appointed as Chair of the Committee
  - Institute is at Univ. of South Florida
- 1986 – Ted Ashford retires
- 1987 – Dwaine Eubanks appointed Director
  - Institute is at Oklahoma State (87-92) and Clemson (92-01)
- 2002 – Tom Holme appointed Director
  - Institute is at UW-Milwaukee (02-08) and Iowa State (08 – present)

The key constituencies

- Practitioners
  – Often motivated by practicality
- Chem Ed Researchers (TUES grant recipients)
  – Often motivated by validation challenges

Exam development

- Chair is named
- Committee is recruited
- First meeting - sets content coverage
- Items are written and collated
- Second meeting - editing items, setting trials
- Trial testing in classes - provides item stats
- Third meeting - look at stats and set exam
- Meetings are held in conjunction with ACS National Meetings (or BCCE)
  – Partial reimbursement to volunteers

Gen Chem Exams

- Full Year Exam (2009, 2011)
- First Term Exam (2005, 2009)
- Second Term Exam (2006, 2010)
- 1st Term Paired Questions (2005)
- 2nd Term Paired Questions (2007)
- Conceptual (1st term, 2nd term, full year)
- Full year - brief exam (2002, 2006)
- All exams carry secure copyright
  - *Released – not published*

Norms and reporting

- Norms are calculated on voluntary return of student performance data
- We have an interactive web site for score reporting for exams that do not yet have enough data to report a norm.
- People often use norm (percentile) to help students who transfer to other programs.

Teachable moments

- Because a sizeable fraction of the Chem Ed community uses (or at least trusts) ACS Exams, the characterization of the exams allows an avenue to educate about assessment issues.

Recently taught topics

- Role of item complexity
- Item characteristic curves
- Item Order Effects
- Answer Order Effects
- Differential Item Functioning (DIF)
- Partial credit / polytomous scoring

Content vs. construct

- Tests demand that students complete tasks
- Each item is a task
- Students need knowledge within the content domain (chemistry)
- Students need knowledge about how to organize their efforts (test taking)
- ***Cast this understanding in terms of item complexity.***

Estimating task complexity elsewhere
Paas & Van Merriënboer's 9-point scale (1994).

1 ~ very, very low mental effort
2 ~ very low mental effort
3 ~ low mental effort
4 ~ lower than average mental effort
5 ~ average mental effort
6 ~ higher than average mental effort
7 ~ high mental effort
8 ~ very high mental effort
9 ~ very, very high mental effort

Objective complexity rubric

Number & relative difficulty/complexity of component concepts or skills needed to master item

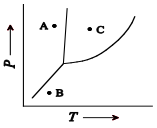
Rating	easy	medium	difficult
1	1		
2	2	1	
3	3 – 4	2	
4	5 – 6	3 – 4	1
5		5 – 6	2
6			3 – 4
7			5 – 6

Concept/skill interactivity

Non-significant	0
Basic	+1
complex	+2

Example: expert objective complexity rating

Consider the phase diagram of a pure compound. Which statement applies?



(A) The path A⟶C represents sublimation.

(B) Following the path A⟶B⟶C, the compound would first liquefy and then vaporize.

(C) If the compound is in state A, continued reduction of the pressure (at constant temperature) will cause it to melt.

(D) None of these statements is correct.

Use the rubic to determine overall objective complexity rating for item:

5 easy concepts (4) + 1 medium concept (2) + complex interactivity (2) = **8** ←

Counting concepts and skills and assign relative difficulty:

- Basic knowlegde of concept of a phase } easy
- Skill: Interpret phase diagram } medium
- Knowledge of phase change related terminology: sublimation, liquefy, vaporize } 3 easy
- Basic knowledge of effects of pressure changes on solids } easy

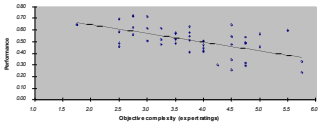
Determine interactivity:

Conceptual/skill interactivity = 2

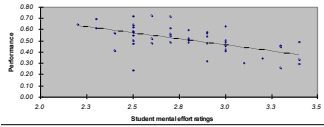
Data for each chemistry exam item

- Performance data (difficulty index)
- Expert-rated objective complexity
- Mental effort (hypothesized to represent the subjective complexity)

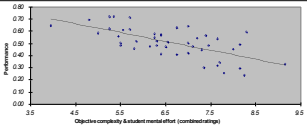
Gen-chem practice exam (45 items)



Multiple R = 0.581
Squared R = 0.338
P-sign = 2.81E -05
y = -0.07492x + 0.79493



Multiple R = 0.570
Squared R = 0.325
P-sign = 4.32E -05
y = -0.21658x + 1.113



Multiple R = 0.663
R squared = 0.437
P-sign = 7.09E -07
y = -0.07302x + 0.991336

Three constructs of task complexity

- Complexity treated as a *psychological experience*.
  - **Subjective complexity**
- Complexity treated as a function of *objective task characteristics*.
  - **Objective complexity**
- Complexity treated as an ***interaction*** between task and person characteristics.

Principle component analysis

Communalities

	Initial	Extraction
error rate	1.000	.700
complexity (rating)	1.000	.629
mental effort (rating)	1.000	.704

Extraction Method: Principal Component Analysis.

Component Matrix[a]

	Component
	1
error rate	.837
complexity (rating)	.793
mental effort (rating)	.839

Extraction Method: Principal Component Analysis.
a. 1 components extracted.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.033	67.768	67.768	2.033	67.768	67.768
2	.537	17.910	85.678			
3	.430	14.322	100.000			

Extraction Method: Principal Component Analysis.

Factor analysis

- Factor Analysis finds a single factor with all PCA loading factors above 0.75
- Hypothesis: This factor represents the ***complexity*** of multiple-choice chemistry items.
- Principal axis factoring and maximum likelihood factoring both reveal a single factor well.

Item design and complexity

- If elements of complexity can result from either content or construct, where do ACS Exams items fall on the complexity scale?
- Focus on conceptual questions.

Construct Complexity in ACS Conceptual Exams

- Does the number or type of elements in conceptual questions correlate to student performance?
- Three ACS Conceptual Exams were analyzed looking at the ***number*** and ***type*** of elements in the question
- A variety of statistical analyses were carried out (software is Stata)

Initial Findings

- First term General Chemistry exam
  - Including a two-part answer is significant (more complex) (p=0.0103) when all answer components are analyzed together
- Second term General Chemistry exam
  - Including a PNOM illustration in the stem is significant (more complex) (p=0.0254) when all stem components are analyzed together
- Second term General Chemistry Paired Questions (Conceptual Items)
  - Model that counts all components shows that construct elements as a whole are statistically significant (p=0.0079)
  - More components – more complex – lower performance
  - *Content is relevant here*
- Some Gen Chem exams show no significant correlations
  - Full year General Chemistry exam and First term General Chemistry Paired Questions (Conceptual Items)
    - No statistically significant findings

Key Preliminary Conclusion

- The construct of conceptual questions on ACS exams is typically not causing students to struggle with questions
- There are no specific question elements that consistently cause student difficulty
- ACS Exams are testing content more than students' ability to navigate the structure of exam questions
  - Recall process – items with poor construct usually do not survive the trial test phase.

What can researchers use?

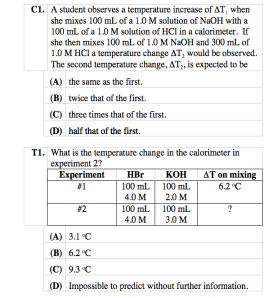
- High quality individual instruments
  - Paired question exams
- A support system for program assessment.
  - Criterion referencing

Paired questions exams

- One each for first semester and second semester General Chemistry.
- 40-item exam, 55 minutes
  - Allows for use of an ACS Exam and a local exam.

Example Pair

- An example of what is meant by a paired question (from trial test – not on released exam.)
- We anticipate this image may be used in publications.



C1. A student observes a temperature increase of ΔT_1 when she mixes 100 mL of a 1.0 M solution of NaOH with a 100 mL of a 1.0 M solution of HCl in a calorimeter. If she then mixes 100 mL of 1.0 M NaOH and 300 mL of 1.0 M HCl a temperature change ΔT_2 would be observed. The second temperature change, ΔT_2, is expected to be

(A) the same as the first.
(B) twice that of the first.
(C) three times that of the first.
(D) half that of the first.

T1. What is the temperature change in the calorimeter in experiment 2?

Experiment	HBr	KOH	ΔT on mixing
#1	100 mL 4.0 M	100 mL 2.0 M	6.2 °C
#2	100 mL 4.0 M	100 mL 3.0 M	?

(A) 3.1 °C
(B) 6.2 °C
(C) 9.3 °C
(D) Impossible to predict without further information.

Looking at 1st Term PQ Exam

- During trial testing it was clear that performance was not unilaterally better for algorithmic items.
- So where are students better?

Coarse Grain Look

Based on 3073 student performances from 12 schools

A 50/50 split!

Average Difficulties:
Conceptual: 0.653
Traditional: 0.598

Topic	Item Pair	Conceptual Difficulty	Conceptual Discrimination	Traditional Difficulty	Traditional Discrimination
Properties of Matter	P1	0.604	0.476	0.712	0.414
	P2	0.870	0.222	0.795	0.419
	P3	0.651	0.489	0.738	0.417
Atoms	A1	0.600	0.559	0.696	0.493
Stoichiometry	ST1	0.519	0.524	0.764	0.454
	ST2	0.812	0.334	0.419	0.547
	ST3	0.851	0.334	0.696	0.455
	ST4	0.456	0.551	0.607	0.680
	ST5	0.460	0.463	0.473	0.715
Gases	G1	0.655	0.479	0.698	0.39
	G2	0.741	0.390	0.715	0.454
	G3	0.609	0.434	0.188	0.338
Solutions	SO1	0.613	0.482	0.636	0.244
	SO2	0.445	0.421	0.404	0.547
Atomic Structure	AS1	0.557	0.423	0.611	0.385
Molecular Structure	MS1	0.611	0.433	0.687	0.490
	MS2	0.611	0.441	0.447	0.562
	MS3	0.866	0.260	0.651	0.488
	MS4	0.804	0.399	0.504	0.325
Thermo-chemistry	T1	0.729	0.432	0.514	0.473

Has something changed?

- The first paired questions exam was released in 1997 in response to evidence that students were lacking conceptual knowledge.
- 12 years later, data suggests responses (e.g., PNOM figures in textbooks) has changed the dynamic.
  - We cannot distinguish if this means students are solving conceptual problems algorithmically.

Criterion referencing for program assessment

- Requires criteria
- At the college level, they don't exist.
- Build a consensus content map.
- Similar to using backward design[1].

1: *Understanding by Design*, Grant P. Wiggins, Jay McTighe

Anchoring Concept

- Use "big ideas" or anchoring concepts to organize content across disciplines.
- Build levels with finer grain size down to the point where exam items are generally written.

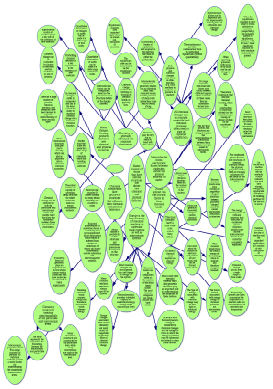
Levels of criteria map

- Level 1 — Anchoring Concept
- Level 2 — Enduring Understanding
- Level 3 — Sub-disciplinary articulation
- Level 4 — Content details

Process for setting map (so far)

- Begin from EMV conference ideas
- Focus Group (Mar08): Level 1 + Level 2
- Workshop (Jul08): Level 2 + Level 3 (General)
- Focus Group (Aug08): Level 2 + Level 3 (Organic)
- Workshop (Mar09): Level 3 + Level 4 (General)
- Focus Group (Aug09): Level 2 + Level 3 (Organic)
- Workshop (Mar10): Alignment (General)
- Focus Group (Mar 10): Level 2 + Level 3 (Physical)
- Focus Group (Jul 10): Level 3 Organic
- Focus Group (Dec 10): Level 3 + Complexity Organic
- Focus Groups (Mar 11): Level 3 (Analytical, Biochem, Physical)

View through 2 levels



Step 2: Alignment

- Look at current items from ACS Exams and align them to Level 3/4
- Process guided by psychometric experts.
- Can include both skills and content
- Ultimately can help define specifications for future ACS Exams.

…and alignment?

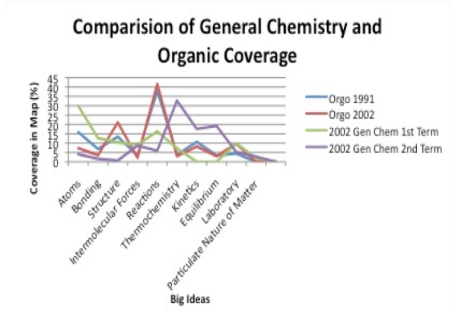
Map Individual Test Items to the Content Map

Item example from General Chemistry exam:
How many neutrons are in ^{19}F?

(A) 9 (C) 19
(B) 10 (D) 28

Big Idea	Enduring Understanding	Subdisciplinary Articulation	Gen Chem Statements
I. Matter consists of atoms that have internal structures that dictate their chemical and physical behavior.	I. Atoms have unique chemical identities based on the number of protons in the nucleus	The number of protons specifies the atomic number which is used as an identifier of the atom/element.	The building blocks of chemical systems, while unseen, represent key definitions for understanding chemistry.
		Atoms may have the same chemical identity but different masses and this explains the existance of isotopes.	Protons and neutrons sum to contribute the mass of an atom.
			Atoms of one element have the same number of protons but can have differing numbers of neutrons. These are called isotopes.
			Measures of mass are made at scales much larger than individual atoms, so elements have atomic masses that are the weighted average of all naturally occuring isotopes.
	II. Electrons play the key role for atoms to bond with other atoms	For a neutral atom there are as many electrons as there are protons, but the electrons can be categorized as core (inner) and valence (outer) electrons.	Valence electrons which determine the properties of elements are correlated with the groups in the periodic table.
		The quantum model of the atom is capable of explaining many observations, and it organizes electrons into "orbitals" that are identified using quantum numbers.	The quantum mechanical model of the atom gives rise to the concept of orbitals.
			Quantum numbers specify these orbitals.

Comparison of gen and org



Summary

- Exams Institute has a stable and trusted process for exam development
  - Process is grass roots
  - Ballpark exam usage is in 15-20% of classes
  - In some places, exams help with articulation issues between 2-year and 4-year schools
- Over the past ~5 years we have begun to engage in more research questions
  - Deliver exams electronically.
  - Enhance validation work.

Acknowledgements

- Kristen Murphy (UWM)
- Karen Knaus (UC-Denver)
- April Zenisky (UMass)
- ISU: Jacob Schroeder (now at Clemson), Heather Caruthers, Mary Emenike, Megan Grunert (now at W. Michigan), Anna Prisacara, Diana Merritt
- Many exam writing committee volunteers

NSF: DUE-0618600, 0717769, 0817409, 0920266

taholme@iastate.edu