Chapter 12

Assessing Conceptual versus Algorithmic Knowledge: Are We Engendering New Myths in Chemical Education?

Thomas Holme*,[1] and Kristen Murphy[2]

[1]Department of Chemistry, Iowa State University, Ames, Iowa 50011
[2]Department of Chemistry and Biochemistry, University of Wisconsin – Milwaukee, Milwaukee, Wisconsin 53201
***E-mail: taholme@iastate.edu**

Studies over the past two decades have emphasized a gap between relatively weak student performance on conceptual items versus traditional items. The ACS Examinations Institute has released a pair of exams for general chemistry in which items are intentionally paired with one conceptual and one traditional item. This paper describes data from statistical analysis of the item pairs, and notes that for these exams, this gap is not evident, as overall performance is better on conceptual items. Possible implications for teaching and for research in Chemistry Education are noted.

Introduction

The prospect that students may learn quantitative problem solving skills within chemistry while not understanding the conceptual basis for the content has been of interest for over 20 years. For example, Nurrenburn and Pickering found that conceptual understanding of stoichiometry lagged behind quantitative understanding (*1*). Subsequently, several groups have confirmed this as well as determined other features. Pickering established (*2*) that performance on conceptual questions in general chemistry was not a predictor of success in organic chemistry. Sawrey showed that difficulties with conceptual items were found for students with both high and low performance on traditional quantitative items (*3*). Nakhleh and coworkers carried out a series of studies that further established the gap between conceptual understandings and algorithmic problem

In Investigating Classroom Myths through Research on Teaching and Learning; Bunce, D.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

solving skills and sought pedagogies to mediate that gap (*4–8*). More recent work by Niaz (*9*) and Cracolice (*10*) continues to identify ways in which conceptual knowledge lags behind algorithmic knowledge in chemistry students. A key component of all of these studies was the use of the paired-question format, where student performance comparisons are drawn from multiple choice item pairs that are designed to provide data about conceptual and algorithmic knowledge separately. The ACS Exams Institute provided a specific tool for this type of assessment in 1997 (*11*) and updated the general chemistry paired questions exams in 2005 and 2007 (*12*, *13*).

The importance of conceptual misunderstandings that were uncovered via this methodology led to a wide range of studies that identified student misconceptions (or alternate conceptions) in a number of content domains of chemistry (*14–16*). In addition to identifying the existence of misconceptions, it is arguable that this data led to changes in the manner in which textbooks presented information about chemistry at the particulate level. Thus, over the past 20 years, since the conceptual/algorithmic gap was first uncovered, there has been both further research and pedagogical responses.

This chapter provides information about student performances on the 2005 Paired-Questions First-Semester General Chemistry Exam (GC05PQF) and the 2007 Paired-Questions Second-Semester General Exam (GC07PQS) that were released by the ACS Exams Institute. These exams have been used nationally for several semesters, and the norm generation process of the Institute (*17*) has allowed for the consideration of item-level analysis of each exam over several thousand student performances. This information can be used to establish two key things. First, the previously identified gap that was abundantly clear 20 years ago (*1–3*) is perhaps not as prominent or as unidirectional today. There are currently item pairs for which student performance on the conceptual item is better than on the algorithmic item. Second, these paired-questions exams show how important the measurement of student performance can be in terms of understanding what students are learning in general chemistry. In particular, exam design plays a critical role in observations about student learning. The current exams show a smaller gap, not only because the student performance database is sensitive to changes in pedagogy and curriculum over the past 20 years, but also because the designers of the exam itself sought an instrument in which the conceptual items and traditional items had more nearly equal performances. This chapter provides abbreviated national normative data for the two exams and also considers guidelines for how the results of these particular assessments are best framed in terms of current pedagogy and learning, and measurement theories.

Exam Development and Structure

The paired questions exams were prepared in a manner similar to the standard procedure for ACS Exams (*17*). The key difference is that not all items in each exam were developed originally for the exam. A number of items were obtained from already released exams. Nonetheless, after all workable item pairs were gleaned from available items on released exams, it was determined that some

In Investigating Classroom Myths through Research on Teaching and Learning; Bunce, D.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

content areas were not adequately covered and specific items or item pairs were developed for the exam. A trial-test phase of the development was undertaken so that student performances could provide statistical data to determine which pairs of items to include on the released exam. This process also led to the development of pairs of items that illustrate item pairs in general chemistry, while not having the security restrictions that forbid the publication of items from ACS Exams. One such pair, from the first-term exam is shown in Figure 1.

C1. As shown in Figure 1, a gas is sealed into a syringe and immersed in a liquid at room temperature (300 K). When the liquid boils the gas in the syringe changes volume as depicted in Figure 2. Assuming no gas leakage and no change in atmospheric pressure, which is the best estimate of the boiling temperature of the liquid?

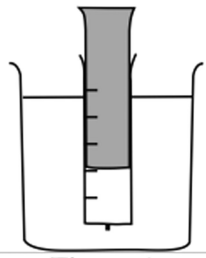

Figure 1

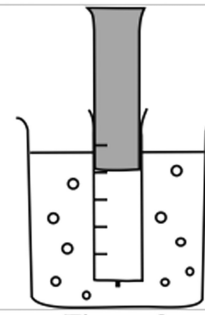

Figure 2

(A)	about 200 K	**(B)**	about 300 K
(C)	about 400 K	**(D)**	about 600 K

T1. A sample of CO_2 occupies 6.00 L at 25 °C and 700. mmHg. What is the volume of the gas at STP?

(A)	4.17 L	**(B)**	5.06 L	**(C)**	6.03 L	**(D)**	22.4 L

Figure 1. Illustration of paired questions. The content area for this pair is gas laws. C1 is classified as conceptual, while T1 is classified as traditional.

Note that this pair was not used on the released exam so it can be reproduced here and referred to as an example of the paired-item format. It is also important to note that this conceptual item does not involve particulate-level representations of chemical systems as were used in the original work (*1–8*). There are conceptual items on the exam that utilize diagrammatic representations, but the construction of conceptual items is notably broader than this construct. Diagrammatic representations include graphical representations of data, such as a phase diagram

In Investigating Classroom Myths through Research on Teaching and Learning; Bunce, D.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

or the image included in Figure 1, but do not include schematic depictions of matter at the particulate level.

Given this basic structure for item pairs, the overall released exams are constructed from 20 pairs. Tables I and II provide the overall structure of the exams in terms of content.

Table I. Content coverage of item pairs for First Term(*18*)

Topic	*Number of item pairs*
Properties of Matter	3 (6 items)
Atoms, Elements and Compounds	1 (2 items)
Gases	3 (6 items)
Stoichiometry	5 (10 items)
Solutions and Concentration	2 (4 items)
Atomic Structure	1 (2 items)
Molecular Structure	4 (8 items)
Thermochemistry	1 (2 items)

Table II. Content Coverage of item pairs for Second Term(*18*)

Topic	*Number of item pairs*
Equilibrium	3 (6 items)
Kinetics	2 (4 items)
Thermodynamics	3 (6 items)
Electrochemistry	3 (6 items)
Solutions	3 (6 items)
Acid/Base chemistry	4 (8 items)
Nuclear chemistry	2 (4 items)

Data that is returned for norm purposes, and reported here, is from students who were allowed 55 minutes (maximum) to complete the released exam. Instructors who purchase the exam are provided with the specific pairings. The conceptual item occurs earlier in the exam than the traditional item, in 39 of the 40 pairs of items. (The exception occurs in kinetics for the second-term exam, and was the result of formatting issues related to the items and their locations on the page.)

In Investigating Classroom Myths through Research on Teaching and Learning; Bunce, D.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

Basic Item Statistics

The Exams Institute calculates item statistics for all released exams based on classical test theory. The item difficulty is assigned as the fraction of students who answer it correctly. This definition results in a counter-intuitive scale, where an items with a higher difficulty value is answered correctly by more students. The second commonly calculated item statistic is the discrimination, calculated by subtracting the fraction of correct answers among the bottom performing students (as determined by their total score on the exam) from the fraction correct among the top performing students. The number of students in the sample for "top" and "bottom" may be varied, and for the values presented in this work the top quarter and bottom quarter of students are used in the calculations. Additional information about overall norms and item statistics have been published elsewhere (*18*).

The item statistics presented for the first-term exam are determined from 3073 student performances from 12 colleges who contributed data voluntarily. For the second-term exam the sample is derived from 3557 students from 9 colleges. Some large courses are included in this data, and in these cases, there are typically multiple sections present, thus there are more than 21 instructors associated with the data included here. Schools that returned data included large research institutions, comprehensive universities, liberal arts colleges and community colleges. The majority of student performances come from large research institutions in part because of the large size of the general chemistry classes at these institutions.

Looking at the data in Tables III and IV there are several key points to consider. First, the number of items for which student performance is better on the conceptual item than the traditional is 10/20 for the first term and 9/20 for the second term exam. Second, the average difficulty for conceptual items is 0.653 and 0.513 respectively for the first term exam and second term exam. The average difficulty for traditional items is 0.598 and 0.538 respectively. Thus, in terms of a mean behavior, conceptual item performance is slightly better on the first term exam and traditional item performance is slightly better on the second term exam. Finally, in topics where there are more than one item-pair it is quite uncommon, over the entire content domain, for performances to exclusively favor one style of item or another. The only examples are in second term thermodynamics and nuclear chemistry where the traditional item shows higher performance in all pairs for that content.

These observations are clearly tied to the test design. Because item choices for the released exams are predicated on performances based on the trial test phase of the exam, the exam committee made conscious choices to have pairs that favor one type of item over another. For example, at the trial exam stage for the first term exam, the average difficulty and discrimination of the conceptual items that were chosen for the released GC05PQF exam were 0.650 and 0.408 respectively. The traditional items selected for the released exam had 0.625 for difficulty and 0.483 for discrimination. Thus, the design of the exam was to have the conceptual and traditional items similar in difficulty. The "predicted" difficulty was quite close for the conceptual items, but the traditional items have now tested slightly more difficult than when they were trial tested. This was also true in the second term

In Investigating Classroom Myths through Research on Teaching and Learning; Bunce, D.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

exam, where the predicted average difficulty based on the trial tests was 0.53 for conceptual items and 0.63 for traditional items (as compared to the observed 0.51 and 0.54.)

Finally, it is also worth noting that the two exams have some apparent structural differences as delineated in Table V. The difference in performance between the conceptual and traditional by item pairs is examined further, identifying a performance gap (by difficulty value) of more than 10% or 20%. It should also be noted that the item pairs included in the >20% category are also explicitly listed in the >10% category.

Table III. Classical Item Analysis for Paired Questions – First Term[†](*18*)

Topic	*Item Pair*	*Conc. Diff.*	*Conc. Disc.*	*Trad. Diff.*	*Trad. Disc.*
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	TH1	0.729	0.432	0.514	0.473

[†] Conc. Diff. = difficulty of conceptual item; Conc. Disc. = discrimination of conceptual item; Trad. Diff. = difficulty of traditional item; Trad. Disc. = discrimination of traditional item.

In the first-term exam, it is much more common for large differences (defined somewhat arbitrarily as greater than a 10% difference in performance) in difficulty for paired items to result from the conceptual item having a much higher difficulty index (i.e. the conceptual item has better performance.) By contrast, in the second-term exam, while the number of pairs with 20% or more performance difference is the same (2 each) the number of traditional items with at least 10% better perfomance is 6 compared to just 2 where the conceptual item shows better performance.

Table IV. Classical Item Analysis for Paired Questions – Second Term†

Topic	*Item Pair*	*Conc. Diff.*	*Conc. Disc*	*Trad. Diff.*	*Trad. Diff.*
Equilibrium	EQ1	0.486	0.442	0.532	0.368
	EQ2	0.495	0.583	0.408	0.486
	EQ3	0.224	0.391	0.445	0.452
Kinetics	K1	0.489	0.433	0.480	0.372
	K2	0.665	0.251	0.622	0.477
Thermodynamics	TD1	0.445	0.375	0.561	0.337
	TD2	0.426	0.427	0.707	0.441
	TD3	0.520	0.429	0.574	0.511
Electrochemistry	EC1	0.403	0.435	0.333	0.382
	EC2	0.513	0.441	0.613	0.475
	EC3	0.576	0.416	0.606	0.470
Solutions	SO1	0.462	0.394	0.417	0.417
	SO2	0.539	0.273	0.453	0.515
	SO3	0.511	0.485	0.749	0.501
Acids/Bases	AB1	0.512	0.629	0.629	0.512
	AB2	0.766	0.412	0.422	0.330
	AB3	0.608	0.449	0.599	0.386
	AB4	0.535	0.381	0.289	0.411
Nuclear	N1	0.561	0.469	0.611	0.443
	N2	0.528	0.475	0.700	0.497

† Conc. Diff. = difficulty of conceptual item; Conc. Disc. = discrimination of conceptual item; Trad. Diff. = difficulty of traditional item; Trad. Disc. = discrimination of traditional item.

In Investigating Classroom Myths through Research on Teaching and Learning; Bunce, D.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

Table V. Item pairs with sizable performance differences

	Higher performance on conceptual items.		*Higher performance on traditional items.*	
	10% higher	*20% higher*	*10% higher*	*20% higher*
First term exam	ST2, ST3, G3, MS2, MS3, MS4, TH1	ST2, G3, MS3, MS4, TH1	P1, ST1, ST4	ST1
Second term exam	AB2, AB4	AB2, AB4	TD1, TD2, EC2, SO3,AB1, N2	TD2, SO3

Discussion and Implications

At this point, the information presented is essentially an empirical observation, predicated on the ability of the Exams Institute to organize both test construction and data collection over nationally relevant student samples. This does not imply, however, that these empirical observations provide no insight into the robustness of theories of learning or assessment related to general chemistry. In particular, it may be possible to infer some hypotheses about how the results summarized here are related to chemistry education from either a research or practice perspective.

The distinction between student performance on traditional chemistry items versus conceptual items has been a key empirical motivation for understanding how students learn chemistry for decades. One hypothesis that may be formulated from the results on this set of exams is that the work of early investigations of this phenomena (*1*–*8*) appears to have had an effect on instruction and student learning. Twenty years after the seminal paper from Nurrenbern and Pickering (*1*) a nationally administered exam can be crafted to measure both aspects of student learning and the resulting student performance is no longer a one-sided measure. As noted, for over 40 item pairs, performance is better on conceptual items in 19 cases and on traditional items in 21 cases. It may be that conceptual understanding gains are more substantial in material commonly covered in the first semester of general chemistry (suggested by the data in Table V), but overly broad generalizations about student conceptual understanding may be risky to make. In essence, care must be exercised to avoid having research results engender new myths about teaching and learning in general chemistry.

There are certainly several caveats that must be acknowledged relative to this conjecture. First, any assessment is the product of the efforts of the writers and carries with it the assumptions (implicit or explicit) they make in its construction. In this case, the committee that constructed the exam had the ability to look at trial test data and choose items that would allow for similar levels of performance, on average, for conceptual and traditional items. The items used in earlier research were generally designed to elicit the misunderstandings that students tend to have, so the expectations of the measurement were different. Nonetheless, it is

In Investigating Classroom Myths through Research on Teaching and Learning; Bunce, D.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

worth noting that even the trial tests were conducted "in the wild", that is, within classroom environments where the test was part of a course.

Second, the sample of instructors who use this exam is small. The number of student performances for this sample is large, but it is possible, perhaps probable, that the instructors who choose to use this particular exam are inclined to include an emphasis of conceptual understanding in their teaching. Otherwise, they would likely use other ACS exams without this same emphasis. It may be that there are classrooms where student performance on the traditional items would be much better than conceptual, because the instructor does not emphasize conceptual understanding. Within test theory (*19*), however, this eventuality would represent a case where the test is utilized outside of its appropriate content domain. Third, general chemistry textbooks from the 1980's when the initial studies were conducted had less emphasis on particulate level, conceptual understanding of chemistry than more recent books. To enumerate this claim, counts were carried out of illustrations that depict the particulate nature of matter (PNM) in a selection of textbooks from the 1980's era, and the current era. Some judgment is required to categorize illustrations. For example, Lewis structures are considered symbolic in this context, rather than PNM illustrations. Orbital illustrations are also not included, in part because they tend to support a different form of pedagogy related to bonding rather than reactivity and in part because there has been relatively little change in the extent of these depictions utilized in texts. Illusrations are designated dynamic if they impart information about either physical or chemical change. The data from this exercise is summarized in Table VI.

The percentage of pages on which any particulate-level images are shown in the older set of books is 5.5%, while in modern texts the value is 30%. The comparison is even more dramatic when the nature of the illustration is considered. Images that imply dynamic characteristics (reactivity) at the particulate level increase from less than 1% to 5%. Dramatically, though not summarized in Table VI, in the 6 older texts, not a single end-of-chapter exercise utilizes a PNM illustration. Across the current textbooks, 236 pages in the 7 texts in the current sample contain at least one problem with such an illustration. This represents over 3% of the total pages. There seems to be little doubt that students today have a better chance of seeing PNM, conceptual depictions than students of 20 years ago.

Finally, students may be learning how to take tests that include conceptual items. From the perspective of learning theory, student test taking can often be understood in terms of which cognitive process is engaged. As categorized by Evans (*20*), there are two systems (System 1 and System 2) that humans access to accomplish a given cognitive task. System 1 tends to be more heuristic (*21*) where System 2 is more analytical and utilizes working memory (*22*). If students have been exposed to tasks that are categorized as conceptual often enough, they may have constructed useful heuristics that allow for facile answers to these questions, regardless of the putative conceptual nature of the item. In the early research in this field (*1–8*), there is little chance that the subjects had such heuristics, because the items were quite novel. Over the past two decades, more test items have emerged that are conceptual based. Student practice at answering these items increases, and test performance improves due to this practice and the heuristic reasoning it induces.

In Investigating Classroom Myths through Research on Teaching and Learning; Bunce, D.;
ACS Symposium Series; American Chemical Society: Washington, DC, 2011.

Table VI. Counts of images of the particulate nature of matter in representative general chemistry textbooks of two eras

Text (Year)	*% of pages with particulate-level images*	*% of pages with "dynamic" particulate images*
Brady and Holum, 1981	7%	1.4%
Chang, 2E, 1984	6%	0.8%
Gillespie, Humphries, Baird and Robinson, 2E, 1989	6%	0.7%
Holtzclaw and Robinson, 8E, 1988	5%	0.5%
McQuarrie and Rock, 2E, 1987	6%	1.0%
Mortimer, 6E, 1986	3%	0.7%
Average – 1980's era	**5.5%**	**0.9%**
Ebbing and Gammon, 9E	33%	6%
Gilbert, Kirss, Foster and Davies, 2E	24%	6%
Kotz, Treichel and Townsend, 7E	31%	6%
Moore, Stanitski and Jurs, 4E	25%	7%
Silberberg, 5E	35%	9%
Tro, 2E	31%	6%
Zumdahl and Zumdahl, 7E	27%	7%
Average – 2010 era	**30%**	**7%**

These, or any other conjectures or hypotheses about the observations for national samples of student performances, do not mitigate the importance of these exams as a tool for instruction and research. It does, however, point to the importance of having a theory base for both instruction and assessment in terms of using these exams. Ideally, test development "in the wild", such as that carried out by the Exams Institute will result in an instrument that has utility for research within a range of possible theory bases. The exams themselves are designed to have value for practicing educators, so long as care is taken to be sure that the content domain covered in the exam matches that of the course in which it is used. In the case of the paired-questions exams, this domain must include the relative emphasis of conceptual understanding.

Finally, it is worth noting that the sample analyzed here is large and representative of a range of instructional strategies. Even if many or most of the instructors value conceptual understanding for their students, there is little doubt that they have varying levels of emphasis on conceptual understanding. Thus, the empirical observation that an assessment can be constructed to span both conceptual and traditional domains is useful in itself. It suggests that the research results of the late 20th century, may have led to instructional changes that

are having measurable improvements in student conceptual understanding in the early 21st century. The results presented here by no means prove this conjecture, but they offer tantalizing evidence that research driven curricular change can be effective.

Finally, the item analysis presented here provides an important benchmark for subsequent usage of these exams in future research. In particular, instructors who implement new pedagogies or other teaching interventions designed to enhance conceptual understanding have a well-characterized tool to measure their innovation.

Acknowledgments

The 2005 First Term General Chemistry Paired Questions and 2007 Second Term General Chemistry Paired Questions exams were created by a committee of excellent educators and researchers led by Diane M. Bunce. We gratefully acknowledge their efforts in the crafting of this exam.

References

1. Nurrenbern, S. C.; Pickering, M. *J. Chem. Educ.* **1987**, *64*, 508–510.
2. Pickering, M. *J. Chem. Educ.* **1990**, *67*, 254–255.
3. Sawrey, B. A. *J. Chem. Educ.* **1990**, *67*, 253–254.
4. Nakhleh, M. B. *J. Chem. Educ.* **1992**, *69*, 191–196.
5. Nakhleh, M. B. *J. Chem. Educ.* **1993**, *70*, 52–55.
6. Nakhleh, M. B.; Mitchell, R. C. *J. Chem. Educ.* **1993**, *70*, 190–192.
7. Zoller, U.; Lubezky, A; Nakhleh, M. B.; Tessler, B.; Dori, Y. J. *J. Chem. Educ.* **1995**, *72*, 987–989.
8. Nakhleh, M. B.; Lowrey, K. A.; Mitchell, R. C. *J. Chem. Educ.* **1996**, *73*, 758–762.
9. Niaz, M. *Int. J. Sci. Educ.* **1995**, *17*, 343–355.
10. Cracolice, M. S.; Deming, J. C.; Ehler, B. *J. Chem. Educ.* **2008**, *85*, 873–878.
11. Eubanks, E. D.; Eubanks, L. P. *General Chemistry Special Exam, First Term*; American Chemical Society, Division of Chemical Education, Examinations Institute: Washington, DC, 1997.
12. Holme, T. A.; Murphy, K. L. *General Chemistry First Term Paired Questions Exam*; American Chemical Society, Division of Chemical Education, Examinations Institute: Washington, DC, 2005.
13. Holme, T. A.; Murphy, K. L. *General Chemistry Second Term Paired Questions Exam*; American Chemical Society, Division of Chemical Education, Examinations Institute: Washington, DC, 2007.
14. Smith, K. J.; Metz, P. A. *J. Chem. Educ.* **1996**, *73*, 233–235.
15. Sanger, M. J.; Greenbowe, T. J *J. Res. Sci. Teach.* **1997**, *34*, 377–398.
16. Raviolo, A. *J. Chem. Educ.* **2001**, *78*, 629–631.
17. Holme, T. A. *J. Chem. Educ.* **2003**, *80*, 594–598.
18. Holme, T. A; Murphy, K. L. *J. Chem. Educ.* **2011**, *88*, 1217−1222.

19. Crocker, L.; Algira, J. *Introduction to Classical and Modern Test Theory*; Holt, Reinhart and Wilson: New York, 1986.
20. Evans, J. S. T. B. *Ann. Rev. Psychol.* **2008**, *59*, 255–278.
21. Chen, S.; Chaiken, S. *Dual-Process Theories in Social Psychology*; Chaiken, S., Trope, Y., Ed.; Guilford: New York, 1999; pp 73−96.
22. Baddeley, A. D.; Hitch, G. J. *The Psychology of Learning and Motivation*; Bower, G. A., Ed.; Academic Press: New York, 1974; Vol. 8, pp 47−90.