Orbital: The Electronic Journal of Chemistry
journal homepage: www.orbital.ufms.br
e-ISSN 1984-6428

| Vol 11 | | No. 1 | | January-March 2019 |

PAPER ON EDUCATION

Psychometric Analysis for Chemical Kinetics Multiple Choice Questions

Luca D'Ottone[a] and Enunuwe Chuckunoye Ochonogor[b]

[a]University of South Africa 1 Preller St, Muckleneuk, Pretoria, 0002, South Africa.
[b]Cape Peninsula University of Technology, 7705, Cape Town South Africa.

Article history: Received: 26 September 2018; revised: 02 January 2019; accepted: 08 January 219. Available online: 30 March 2019. DOI: http://dx.doi.org/10.17807/orbital.v11i1.1323

Abstract:

The aim of this study is to research the effectiveness of MCQs as a testing tool for undergraduate chemical kinetics. This study analyses psychometric indicators of a 30 multiple choice questions (MCQs) test focused on chemical kinetics, and compares those indicators with the ones calculated for commonly accepted American Chemical Society (A.C.S.) inorganic chemistry test. The study group consists of first and second year General and Inorganic Chemistry students of a major college in the Southeast region of the United States of America (N=68). *Quasi*-experimental design was used for this research. After authorization of the Institutional Research Board of the institution was secured, a group of 104 students was ask to participate. Out of these, only 68 decided to move forward and participate in the testing. The participants, after being exposed to different study materials were tested on their understanding of chemical kinetics with a 30 MCQs test. The psychometric indicators for the 2015 General Chemistry ACS standardized test were calculated from information available on the official web site of the ACS Examination Institute (EI). The MCQs test built on a battery of 30 questions of the present study demonstrate to have an accepted difficulty level (p=51.08), good internal consistency (KR-20= 0.76), but low discriminatory power (D=0.14). The results of the MCQs test built for this study are generally consistent with the benchmark that can be inferred from the ACS EI data, and underscore the fact that undergraduate chemical kinetics can be effectively tested *via* a battery of MCQs.

Keywords: chemical kinetics; multiple choice questions; psychometric Indicators

1. Introduction

Multiple Choice Questions (MCQ) are a widely used testing tool in general education. Chemistry is no exception: the Chemical Education Examination Institute (EI) of the American Chemical Society (ACS) in cooperation with the University of Milwaukee recently launched a revised and modernized version of the Organic Chemistry (first and second semester) and General Chemistry (first and second semester) standardized examination tests built on MCQs. The goal of the present study is to determine the feasibility of testing college chemistry with MCQs, and to explore what are realistic expectations for expectations for the scores for a college chemistry exam built on MCQs. This goal is achieved by performing an independent field investigation on first and second year college chemistry students, and comparing the results of the field investigation with a generally accepted reference such as the ACS EI General Chemistry standardized examination.

Chemistry educators, just like academics of other allied sciences [1, 2], use as a reference a set of generally accepted recommendations for the development of MCQ tests [3]. These recommendations include, amongst others, the following:

- Writing questions after clear learning objectives;
- Developing a pool of questions ahead of time, and making the selection just before the test date;
- Utilize proper grammar, punctuation, spelling, and nomenclature;
- Avoiding overly wordy stems;

*Corresponding author. E-mail: ldottone@mdc.edu

- Avoiding "*all of the above*" or "*none of the above*" kind of answers; and
- Avoiding two tiers questions.

Because of the relatively high degree of abstraction required, MCQs directed to learning objectives related to chemical kinetics should be at the application level in the Bloom taxonomy [4, 5]. In addition to the guidelines above, Campbell [6] suggests different strategies to discourage students from guessing at their general chemistry MCQs test. These strategies, while not universally adopted, include, for example, giving at least partial credits for questions left blank. Despite the convenience of MCQs, a preliminary survey of the ACS Examination Institute discussed by Brandried, Reed, and Holme, reveals that still most chemistry instructors do not rely uniquely on MCQ when they verify the level of maturity of first and second year college students [7]. This preference, of not entirely relying on MCQs, underscores the scepticism, often not openly manifested, of most faculty in reducing higher thinking skills into bullets. Stanger-Hall presented evidences supporting the scepticism of the instructors that do not adopt MCQs as their unique tool to assess students [8]. While MCQs are very popular and extremely practical, they do not support the development of critical thinking and it may create an illusory feeling of competence. This is due to students thinking they know the subject matter as they are able to answer correctly a battery of questions, on the other hand the false confidence is quickly crushed when students are challenged with real tasks such as performing an experiment in a laboratory setting.

The ACS standardized tests are not developed in a void [9]. They were developed allowing chemistry instructors throughout the fifty states to participate in the MCQs' drafting process almost embedding the practice of science, into the standardized exam [10]. The ACS EI serves the role of a lead by providing an Anchoring Concept Content Map (ACCM) that outlines the topics to be tested and the level of complexity required for a question to qualify [11]. The ACCM covers in detail ten key concepts of general and inorganic chemistry including:

I. Matter consists of atoms that have internal structure that their chemical and physical behaviour;

II. Bonding: atoms interact via electrostatic forces to form chemical bonds;

III. Structure and function: Chemical compounds have geometric structures that influence their chemical and physical behaviour;

IV. Intermolecular Interactions: Intermolecular forces-electrostatic forces between molecules- dictate the physical behaviour of matter;

V. Chemical reactions: matter changes, forming products that have new chemical and physical properties;

VI. Energy and thermodynamics: Energy is the key currency of chemical reactions in molecular scale system as well as in macroscopic systems;

VII. Kinetics: chemical changes have a time scale over which they occur;

VIII. Equilibrium: all chemical changes are, in principle, reversible; chemical processes often reach dynamic equilibrium;

IX. Experiments, measurements, and data: Chemistry is generally advanced via experimental observation; and

X. Visualization: chemistry constructs meaning interchangeably at the particulate and macroscopic level.

These key concepts were first identified at the 2008 ACS National meetings and were further explored and refined over the course of four years. Currently, the EI considers submissions One common thing that the ACCM has in common with the experimental observation of the present study is that is not tided to one particular textbook, rather is based on outcome-based learning objectives, departing, in a way, from the common predicament, where students are literally trained to use a textbook and resolve the problems associated with that particular textbook provided by the publisher. Because of the grassroots process in developing the examinations, it is safe to say, that even if the ability of testing higher thinking skills through MCQs is still under debate, testing college chemistry by MCQs is a generally accepted practice.

Chemical Kinetics is a specialty of the chemical arts focused on providing a mathematical description of the elementary steps

occurring during a chemical transformation [12]. While there is a general agreement on the formalism used to describe the different temporal profiles, in many occurrences the very own nature of the elementary processes taking place for each reaction remains, at least in part, under investigation. Because of the relative complexity of using a mathematical operator, such as a differential equation, to describe a chemical process, chemical kinetics is often considered a challenging topic by students and teachers alike [13]. The extensive review by Bain & Towns put in evidence the different techniques used to test, mainly Turkish, students in chemical kinetics [13]. These techniques include multiple choice questions, but also include open ended questions, five and seven points Likert scale questions, and two tiers questions. Geicos, Salta, and Koinis analyse the difficulties in teaching and learning chemical kinetics in Greek high school students [14]. Their theoretical framework is very ample, including key studies by Gilbert and Justi [15-17]. On the other hand, rather than blaming the challenges associated with teaching and learning chemical kinetics on the high level of math require to understand this discipline, they reach a different conclusion: they suggest that the inconsistencies in the teaching material creates even more confusion than the difficulties embedded in the subject matter itself. In view of the discordances described above there is the need to address the question whether it is possible to effectively test students' understanding of chemical kinetics in a uniform way with MCQs written in consideration of the recommendation by Towns [3].

Supporting students in a virtual context when ask to simulate chemical experiments with a mathematical processor such as TENUA [18, 19], may be identified as a unique form of servant leadership [20]. Under the servant leadership theory, the leader, in an educational context the instructor, is a subject matter expert and enjoys helping others and leading to their ultimate goal of fulfilling their academic requirements. While servant leadership is a generally positive leadership style fitting into the educational context of constructivism, it also requires a substantial amount of efforts by the part of the faculty [21]. On the other hand, because support is provided unconditionally, when properly enacted it becomes very effective. Convincing college students to use a software program and manipulated data on a spreadsheet it may be challenging. On the other hand, if students are provided constant support and redirected toward that goal, with simple and understandable instruction, they usually manage to do it.

The Attention Relevance Confidence and Satisfaction (ARCS) model of motivational design theories also support the present study [22]. By engaging students to work on their own in a virtual environment, the instructor can stimulate the perceptual arousal needed to gain the full attention of the student. The instructor only supports the students' projects by leading them to success, without directly instructing them or performing the experiment for them. In this way the presumption is created that students are self-directed unless they encounter a holdup, in which case the instructor is consulted to troubleshoot. As students can choose when and how to work in an asynchronous environment, with the idea that they will have to report their results in a formal meeting only in a few weeks, thus building up self-confidence, and ultimately satisfaction about their own work as they see their posts acknowledged in real time by the instructor.

This study was designed within a combination of the servant leadership theory combined with the ARCS model of motivation. The elements of expecting students' performance, typical of the transactional leadership, and surprise, as in the ARCS model, are combined in a unique virtual environment. In a real, live, class a student could be intimidated by being challenged in front of other participant, while in a virtual environment where the performers are publicly recognized and the low performers are not weeded out, the participant enjoy that feeling of anonymity that make feel secure. Live class interaction, must be alternated to virtual interaction, as the latter ones only provide support for the learner: while the bulk of the work is done in a real classroom.

Because of the specificity of the challenges posed to the students, the instructor bears the burden of dissecting the material in simple components that can be tested within the simplicity of a text message: on the other hand, this breaking down the material into simple building blocks, not only make it suitable for a challenge-reward system, but also demystify the complexity of longer chapters that often keeps the students away from the books.

2. Results and Discussion

Out of the 68 students that participated in the study 46 composed the experiment group and 22 composed the control group. 39 participants were female students, 29 participants were male students. The average raw score for the experiment group resulted 21.82 (or 62.34 %) associated to a standard deviation of 4.22. The experiment group mode was 21 and the experiment group median 23. The average score for the control group was 23 associated with a standard deviation of 3.36, the mode of the control group was 25, and the median was 23.5. Table 1 reports the raw scores and the relative frequency for the scores of the experimental and control groups. A Shapiro-Wilk normality test [23] performed on both distributions reveals that, while the distribution of the scores of the control group is normal, the distribution of the scores of the experiment group is not (W=0.84). Therefore, the *t*-student test may not be the most appropriate tool to analyse the data and the Mann-Whitney for independent samples was used instead. For a number of data n>25 the Mann-Whitney [24] test generated a U value of U=449.5 indicating that for p<0.05 (two tail test) there was no significant difference between the two distributions. These conclusions are consistent with the analysis in D'Ottone and Ochonogor [25]: therefore, both sets of data were combined in one homogeneous distribution to study the effectiveness of the questions. The mean raw score for the overall distribution including all the 68 individuals was 22.20 associated with a standard deviation of 3.95.

On the complete dataset, composed by 68 individuals, an extensive analysis, inspired by the one performed by Holme and Murphy [26], was performed. This analysis included the determination of the difficulty index *p*, the determination of the discriminatory index D, the Kuder-Richardson-20, and the Kuder-Richardson-21 parameters. To achieve this, the scores were then divided into three groups: the high achiever (H), the top 73% that scored at least 25 of the answers correctly, and the low achievers (L). The bottom 27% that scored 21 or less the answers correctly and the medium performing group (M) defined as the group that answered more than 21 but less than 25 questions correctly. The high achiever group consisted of 23 students, while the low achievers group consisted of 25 students. As in Islam & Usmani [27] the difficulty index, or facility for each question was calculated with the formula:

$$p = \frac{(H+L)}{N\ X\ 100} \qquad (1)$$

where H is the number of students that answered the question correctly in the high achievers group, L is the number of students that answered the question correctly in the low achiever group, and N is the total number of students. The difficulty index *p* has the format of a percentage ranging from 0, for questions that no one answered correctly, to 100 for questions that everyone answered correctly. Values of *p* in between 40 and 60% are optimal, while values below 30% and above 70% indicate that the question was too difficult, or too easy respectively. The indicators of centre related to the difficulty index *p* for the test examined in the present study were mean=51.08, median= 52.94, and mode=67.67, while the indicators of dispersion were standard deviation= 14.57, and range= 64.71. The Discriminatory index D was calculated according:

$$D = \frac{(H-L)\ X2}{N} \qquad (2)$$

The discriminatory index D varies between -1 and 1 with values above 0.2 being acceptable: indicating a high probability to be guessed correctly by the high achieving group, and values below that indication poor ability to discriminate. The values for the difficulty index *p* and for the discriminatory index D are reported in Table 2. The indicators of centre related to the discriminatory index D for the test examined in the present study were mean=0.14, median= 0.15, and mode=0.00, while the indicators of dispersion were standard deviation= 0.14, and range= 0.50. From the combination of indicators of centre and indicators of dispersion above it is possible to infer that the questions developed for this study gave a reliable indication as to whether the students understood the material or not (0.6>p=0.51>0.4) but shown almost no discriminating power (D=0.14<0.2). The low discriminating power of the set of multiple choice questions developed for the present investigation may be due to the fact that students answered all questions across the board, without necessarily finding any specific question either "easy" or "difficult". This low discriminatory power, may be due, at least in part, to the fact that

students were exposed to different reference material: therefore developing some aspects more than other ones depending on the group in which they participated, either experiment or control. On the other hand, the combinations of *p* and D values indicate that Questions (Qs) 4, 5, 6, 8, 12, 15, 19, 20, 21, 22, 23, 24, 25, 26, 27, and 28 were in fact both a good indication of the understanding of the learning objectives, and a good discriminant between the high performers (H) and the low performers (L). On the other hand, Qs 1, 3, 9 and 30 were eventually either too easy or poorly discriminants between H and L. Other confounding factors included:

- Adherence of the teaching material with the learning objectives,
- The use of different teaching material,
- Attitude of the students toward the subjects, and
- Others.

With respect to the adherence of the teaching material, a correlation analysis performed on all thirty items reveals that some questions were eventually out of line. For example, Q1 was answered correctly by *all* students: this may indicate that it was too easy. Q10 was answered correctly only by a minority of the group: this may be an indication that Q10, while clear to the revising faculty, was not clear to the students.

Table 1. Raw scores, out of thirty questions, and the relative frequency of the correct answers for experimental and control groups of the present study.

Experiment Group		Control Group	
Raw Score	**Frequency**	**Raw Score**	**Frequency**
8	1	17	1
11	1	18	2
12	1	20	3
14	1	21	2
16	2	22	2
17	1	23	1
19	1	24	1
20	2	25	6
21	7	26	1
22	3	27	2
23	6	30	1
24	7		
25	7		
26	5		
27	1		
Total	46	Total	22

Table 2. Numerical values for the difficulty index *p* and for the discriminatory index D of the present study.

Item number	*p*	D	Item number	*p*	D
1	70.59	-0.06	16	67.65	0.00
2	67.65	0.00	17	63.24	0.09
3	69.12	-0.03	18	63.24	0.09
4	42.65	0.32	19	58.82	0.18
5	52.94	0.18	20	60.29	0.15
6	55.88	0.24	21	42.65	0.38
7	35.29	0.06	22	52.94	0.24
8	39.71	0.26	23	44.12	0.41
9	50.00	-0.06	24	54.41	0.26
10	5.88	0.00	25	36.76	0.21
11	67.65	0.00	26	51.47	0.21
12	36.76	0.44	27	29.41	0.18
13	64.71	0.06	28	30.88	0.15
14	64.71	0.06	29	44.12	0.00
15	57.35	0.21	30	51.47	-0.03

The use of different teaching material may have affected, to some degree the outcome of this study. While, in fact, we found no statistically significant difference between the experiment and the control group, a punctual analysis reveals that Q4 and Q25 generated the highest gap, respectively 23% and -27%, in between the groups. Table 3 reports the number of correct answers normalized to a percentage for both the experiment and the control group, showing, implicitly, these gaps. Other confounding factors, may have been students cheating, either among themselves, or independently with cheat sheets, or via telecommunication devices such as smartphones. While every attempt in fact was made to minimize cheating, this is always a

possible source of bias in psychometric testing. The degree of internal reliability was calculated for each question with the Kuder-Richarson 20 (KR-20) formula:

$$[KR-20] = [1 - \frac{\sum_{i=1}^{K} pq}{\sigma_X^2}] \frac{K}{K-1} \qquad (3)$$

Where K is the test item numbered from *i*=1 to K=30, *p* is the proportion of correct answer for item 1, *q* is the proportion of incorrect answers for item *i*, and σ^2 is the variance for the distribution [28]. The [KR-20] scores varies between 0 and 1: these values are also referred as alpha Cronbach coefficient [29]. Values closer to 0 indicating some inconsistencies in the level of difficulty of the test under examination. Values closer to 1 indicating a high level of consistency [30, 31]. The numerical value calculated for the internal consistency resulted to be in the acceptable range [KR-20] = 0.76, suggesting that the overall structure of the test was internally consistent. The information provided by the EI on their web site was not sufficient to estimate the [KR-20] for the 2015 1st term General Chemistry test.

Table 3. Number of correct answers normalized to a percentage for both the experiment and the control group.

Item number	Correct Answers (in percentage)		Item number	Correct Answers (in percentage)	
(1-15)	Experiment Group (n=46)	Control Group (n=22)	(16-30)	Experiment Group (n=46)	Control Group (n=22)
1	100.00	100.00	16	97.83	95.45
2	95.65	100.00	17	89.13	95.45
3	97.83	100.00	18	91.30	95.45
4	73.91	50.00	19	86.96	86.36
5	76.09	86.36	20	84.78	90.91
6	78.26	100.00	21	69.57	59.09
7	36.96	45.45	22	73.91	90.91
8	54.35	50.00	23	69.57	77.27
9	69.57	63.64	24	78.26	86.36
10	4.35	9.09	25	46.65	72.73
11	93.48	100.00	26	78.26	81.82
12	47.83	54.55	27	36.96	59.09
13	91.30	90.91	28	32.61	59.09
14	91.30	95.45	29	73.91	54.55
15	82.61	86.36	30	80.43	63.64

The ACS EI does not provide, at least on their web site, sufficient information to calculate the KR-20 for their test. On the other hand, ACS EI does provide the numerical value for the Kuder-Richardson 21[KR-21]. [KR-21] is another estimate of internal reliability, based on the assumption that there is a high correlation between the different questions. [KR-21] is calculated according to equation (4) as:

$$[KR-21] = [\frac{n}{(n-1)} * [1 - \left(\frac{M*(n-M)}{n*s^2}\right)]] \quad (4)$$

In equation (4) n is the number of items (n=30), M is the average or mean score of the test (M=22.20) and σ is the standard deviation for the distribution (s=4.95) [32, 33]. [KR-21] varies between 0 and 1, values closer to 0 indication a poor internal consistency, and values closer to 1 constituting an indication of higher internal consistency. [KR-20] and [KR-21] are not necessarily directed to measure the same thing. While they are both indicator of the inter-item consistency of the questions or a test, [KR-20] is generally considered more reliable, while [KR-21] is an adaptation of the [KR-20] for a set of questions of same or similar difficulty. The [KR-21] values are numerically lower than the [KR-20] values. In an attempt of comparing the inter-item consistency of the items developed for the present study, an estimate of the [KR-21] was also developed for the MCQs of the present investigation. As described above, no effort was made to ensure the homogeneity of the MCQs in this study, rather they were draw to different level of difficulty, as described in Table 5. Originally, the numerical value of [KR-21] for the test object of the present study is 0.65 suggesting a mediocre consistency between the items developed for this study. That was in some way expected as the MCQs of the present study were directed to a

variety of topics and level of difficulty to cover the whole material discussed both in the experiment and in the control group. Two reasons of concerns then remain with regards to the way that the ACS EI express their results: (1) It is not clear why an indicator of inter-item consistency built in the assumption that all questions have the same or similar level of difficulty; (2) if the [KR-21] of the ACS 2015 EI examination is 0.9, the calculated [KR-20] value should be comprised in between 0.9 and 1: a range of quasi-perfection that is outside the reach of most test. These concerns were not addressed in the present study, rather they will be subject of further analysis.

Table 4. Blue print of the MCQs test according to Bloom's (1965) Taxonomy of Educational Objectives subdivided according to the three main topics of the exam: (1) general understanding of chemical kinetics, (2) order of reaction, and (3) *pseudo*-first order approximation.

	Lower Order Questions			Higher Order Questions			
Topic	**Knowledge**	**Comprehension**	**Application**	**Analysis**	**Synthesis**	**Evaluation**	**Total**
(1)	2	2	1	1	1	9	10
%	6.7%	6.7%	3.3%	3.3%	3.3%	10.0%	33.33%
(2)	2	2	2	2	1	1	10
%	6.7%	6.7%	6.7%	6.7%	3.3%	3.3%	33.33%
(3)	1	2	2	2	2	1	10
%	3.3%	6.7%	6.7%	6.7%	6.7%	3.3%	33.33%
Total	5	6	5	5	6	5	30
%	16.7%	20.0%	16.7%	16.7%	20.0%	16.7%	100%

The potential for gender bias was also analyzed, as suggested by Stanger-Hall [8], was also analyzed. Out of the 68 individuals participating in the study 39 were female and 29 were male. A *t*-student test [34] of the score distributions for female and male students indicated a mean of 21.64 associated with a standard deviation of 4.77 for the females, and a mean of 22.67 associated with a standard deviation of 3.34 for the males. The two-tailed *t*-student variable with a $p<0.05$ confidence interval resulted 0.3051 lower than the critical value of 1.0338: therefore, no statistically significant difference was observed between the scores of the females and the males.

Comparing these results with the 2015 1st term General Chemistry (GC) form of the ACS EI is only possible in part, since ACS publishes a limited report on the outcomes of these exams. Out of a 70 MCQs test the mean score is 39.7 (or 56.71%), the median is 40, the standard deviation is 12.4, and the KR-21 reliability is 0.90. While it is known that [KR-20] is a number between 0 and 1, and it is higher than [KR-21] the exact value for the [KR-20] of the 2015 1st term GC ACS test cannot be calculated with the data provided by the EI. Since the ACS EI data are normalized the difficulty index *p* must equal 54, and the discriminatory power D must be 0. A comparison of the respective psychometric indicators for the test developed for the present study and the 2015 ACS EI GC test is summarized in Table 5. This comparison indicates strong similarities for most numerical values implicitly suggesting, that MCQs are an effective and consistent tool for testing undergraduate chemical kinetics. Figure 1 is the histogram built by comparing the scores earned by students at the 2015 ACS General Chemistry test with the scores earned by the students during the course of the present investigation. The Y-axis is the relative frequency, and the X-axis is the *z*-score calculated according:

$$z=(a_i-\mu)/\sigma \quad (5)$$

where a is the value of a specific *i* exam score, μ is the average or mean exam score for the entire population, and σ is the standard deviation associated to μ.Initially the determination that test subjects studying chemical kinetics with the traditional book and lecture approach would earn scores statistically indistinguishable from subjects that acquired the same knowledge by performed computer based simulations was somewhat unexpected and troublesome [25]. Common knowledge would suggest that: if students learn the material from different sources their average scores on a test may differ. On the other hand, both the 2015 ACS General Chemistry test observed on more than seven thousands students

in 47 different institutions, each following its own curriculum and free to adopt its own book, and the present test, feature a normal distribution of the scores. The shape of the distribution of the scores of the 2015 ACS General Chemistry test almost overlaps to the normal curve, while the shape of the distribution of the scores of the present study is less regular, featuring two peaks respectively for values of z=0 and 1. The peaks at z=0 and 1 are not associated with different modes, having ruled out analytically the possibility of a multimodal distribution [25, this study]. The less than perfect shape of the distribution of the scores of the present investigation may be due, at least in part to the smaller size of the sample. No information is known about the internal distributions of the data for the ACS 2015 General Chemistry test. If one then consider that both distributions behave normally, it can be argued that for test built around outcome-based learning objectives the source of the material is not relevant: this is in contrast to some degree to the predicament, where students are literally told to use a textbook and resolve the problems associated with that particular textbook provided by the publisher.

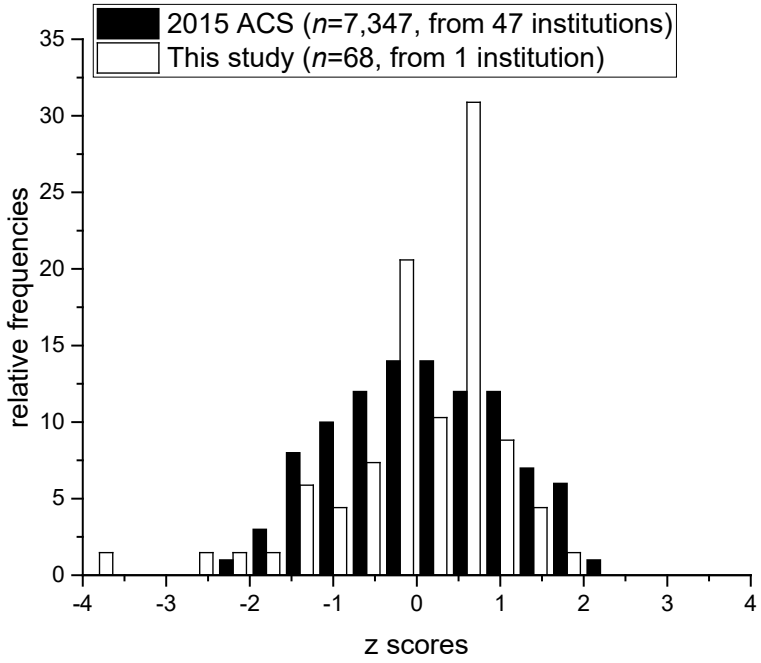

Figure 1. Histogram comparing the relative frequency distributions of the scores earned at the 2015 ACS General Chemistry test, and at the test developed for the present investigation.

Table 5. Comparison of psychometric indicators between the test developed for the present study and the 2015 ACS EI General Chemistry test.

	Test of the Present Study	2015 ACS EI GC Test
mean ± std. dev.(*)	62.34 ± 4.22	56.71 ± 1.82
difficulty (*p*)	51.08	54.00
Discriminatory Power (D)	0.14	0.00
KR-20	0.76	NA
KR-21	0.65	0.90

(*) in percentage points

2. Material and Methods

The present study was conducted at a major college in the Southeast United States in the Academic years 2015-2016 and 2016-2017. Permission was requested to and granted by the Institutional Research Board (I.R.B.) of the college. First and second years college chemistry students were asked to participate in the study at their own discretion. Those who choose to participate where explained the procedures to follow and were asked to sign off an informed consent. Out of a total of 104 students that were asked to participate 68 agreed and took part in the experiment. The students were divided into two groups, to test for possible bias related to the teaching material, a control group assigned to work with the class textbook [35], an experimental

group assigned to work with a classic physical chemistry textbook [36] paired with other literature as described in D'Ottone & Ochonogor [25].

A battery of 30 MCQs, each consisting of a stem, three distractors, and one key, was developed in accordance with the recommendation by Town [3]. Being designed to test the ability of students to reach conclusions about chemical kinetics following different paths, represented either by the traditional book and lecture approach on one side, or by a computer-based simulation approach on the other side, the MCQs developed for the present investigation were outcome based rather than testing a specific detail. The MCQs were then each reviewed by a peer college faculty and by an external advisor at the Ph.D. level from a research university. Table 4 presents the breakdown of the questions according to the Bloom taxonomy [4] as explained by Kim et al. [5]. Figure 2 is typical MCQ used for the present study, consisting of a stem, three distractors, and a key. The battery of test was administered three to six weeks after having introduced the material in class.

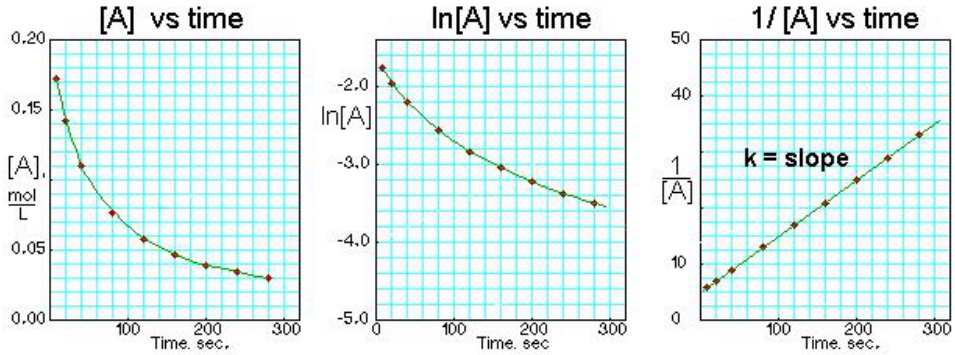

Figure 2. A typical Multiple Choices Question, consisting of a stem, three distractors, and a key, developed for the present study. The textbook reference page, as shown in the figure, was included in the forms submitted for peer review examination and to the external advisor. Q(xx) The following plot represents: a) A zeroth order reaction's profile; b) A first order reaction's profile; c) A second order reaction's profile; d) It cannot be determined; Correct Answer (c) p 591.

The ACS EI tests, that were used as a benchmark for comparison in the development of the present study, are copyrighted and confidential. Therefore, it is no possible to share with the public their contents. On the other hand, because of the prominent position they occupy in the chemical education framework it is important to factor them into any study of this kind. The ACS EI publishes the normalized overall results of their exams in their official web site together with important statistical indicators. The results of the 2015 General Chemistry test were taken as a comparison. The 2015 ACS EI General Chemistry test was a 70 MCQs test covering all the subjects of the first and second year General and Inorganic Chemistry course, including chemical kinetics. These results were collected out of a population of 7,347 students out of 47 institutions, including the institution where this study was performed.

4. Conclusions

Writing an assessment test, is a complex task. In the course of the present study a 30 MCQs test was created to assess the understanding of chemical kinetics by first and second year college students from a major college in the Southeast United States. A sample of 68 students were divided into an experiment group composed by 46 individuals, and a control group composed by 22 individuals. The two groups were exposed to different teaching materials, but the distribution of their overall scores was not statistically different.

The test itself was analyzed its difficulty index *p*, its discriminatory index D, and its internal consistency with the KR-20 formula. Based on the answers observed in this study, the test showed an acceptable difficulty p=51.08±14.57, a good internal consistency KR-20= 0.76, an acceptable KR-21=0.65 value, and a marginal discriminatory power D=0.14±0.14. These results were remarkably consistent with psychometric indicators that can be inferred for the ACS EI 2015 General Chemistry exam that was used as a

benchmark for comparison. What it can be inferred from this experience is:

1. Instructors can develop their own MCQs assessment effectively, as the psychometrics indicators for the in-house test seems consistent with the ACS benchmarks, as summarized in Table 5;

2. Remarkably, the expected average score of MCQ chemistry tests lays in the fifties, therefore the grading scale should be realistically adjusted to these values;

3. The average scores are independent from the study material: both for the test of the present study and for the ACS EI test a variety of study material was used to prepare the students. This may be due in part to the increased availability of information in the digital era. Other factors may include students' prior knowledge of the subject matter, their reasoning skills, and time and efforts dedicated to find study materials elsewhere such as on the internet. This information could be developed more into a completely different study taking in consideration the cost of textbooks and the barrier it creates in the students' educational journey;

4. There is no gender bias, with limited reference to the test of the present study no gender bias was detected by commonly accepted statistical techniques.

Even in the most accurately prepared investigation there may be confounding factors. Confounding factors could include cheating, guessing, using devices such as smartphone to circumvent the test area security and others. More research is needed to determine what are realistic average test scores for chemistry MCQs tests and how they are actually adjusted to the high expectations sets forth by faculties in their Syllabi. Another aspect of the present investigation that could be developed is weather the relatively low scores achieved by students both at the ACS standardize test in exam, and in the MCQ test of the present study are so low because of their disconnect with a specific textbook, rather focusing on outcome based learning objectives.

Acknowledgments

Luca D'Ottone would like to thanks Dean Mark Kraus, Ph.D. and Professor Yaelis Rivas, Ph.D. for the support provided in obtaining the IRB approval.

References and Notes

[1] Considine, J.; Botti, M.; Thomas, S. *Collegian* **2005**, *12*, 19. [Crossref]

[2] Pugh, D.; De Champlain, A.; Gierl, M.; Lai, H.; Touchie, C. *Medical Teacher* **2016**, *38*, 838. [Crossref]

[3] Towns, M. H. *J. Chem. Educ.* **2014**, *91*, 1426. [Crossref]

[4] Bloom, B. S. Taxonomy of Educational Objectives, vol. 1. New York: McKay, 1965.

[5] Kim, M. K.; Patel, R. A.; Uchizono, J. A.; Beck, Lam. *J. Pharm. Educ.* **2012**, *76*, 114. [Crossref]

[6] Campbell, M. L. *J. Chem. Educ.* **2015**, *92*, 1194. [Crossref]

[7] Brandriedt, A.; Reed, J. J.; Holme, T. *J. Chem. Educ.* **2015**, *92*, 1798. [Crossref]

[8] Stanger-Hall, K F. *CBE-Life Sciences Education* **2012**, *11*, 294. [Crossref]

[9] Murphy, K.; Holme, T.; Zenisky, A.; Caruthers, H.; Knaus, K. *J. Chem. Educ.* **2012**, *89*, 715. [Crossref]

[10] Reed, J. J.; Brandriet, A. R.; Holme, T. A. *J. Chem. Educ.* **2016**, *94*, 3. [Crossref]

[11] Holme, T.; Murphy, K. *J. Chem. Educ.* **2012**, *89*, 721. [Crossref]

[12] Pilling, M. J.; Seakins, P. W. Reaction Kinetics. Oxford, U.K.: Oxford University Press, 1996.

[13] Bain, K.; Towns, M. H. *Chem. Educ. Res. Pract.* **2016**, *17*, 246. [Crossref]

[14] Geicos, T.; Salta, K.; Koinis, S. *Chem. Educ. Res. Pract.* **2017**, *18*, 151. [Crossref]

[15] Justi, R. In Chemical education: toward research-based practice. Gilbert, J. K.; De Jong, O.; Justi, R.; Treagust, D. F.; Van Driel, J. H., eds. Dordrecht: Springer, 2002, chapter 13.

[16] Justi, R.; Gilbert, J. K. *Sci. Educ.* **1999**, *8*, 163. [Crossref]

[17] Justi, R.; Gilbert, J.K. *Sci. Educ.* **1999**, *8*, 287. [Crossref]

[18] Barshop, B. A.; Wrenn, R. F.; Frieden, C. *Anal. Biochem.* **1983**, *130*, 134. [Crossref]

[19] Wachsstock, D. H.; Pollard, T. D. *Biophys. J.* **1994**, *67*, 1260. [Crossref]

[20] Greenleaf, R. K. The servant as leader. Notre Dame: University of Notre Dame Press, 1997.

[21] Phipps, K. A. *Journal of Leadership Education* **2010**, *9*, 151. [Link]

[22] Keller, J. M.; Motivational design for learning and performance: The ARCS model approach, Berlin: Springer Science & Business Media, 2009.

[23] D'Ottone, L.; Ochonogor, C. E. *Orbital: Electron. J. Chem.* **2017**, *9*, 299. [Crossref]

[24] Shapiro, S. S.; Wilk, M. B. *Biometrika* **1965**, *52*, 591. [Crossref]

[25] Mann, H. B.; Whitney, D. R. *The annals of mathematical statistics* **1947**, *18*, 50. [Link]

[26] Holme, T.; Murphy, K. *J. Chem. Educ.* **2011**, *88,* 1217. [Crossref]

[27] Islam, Z.; Usmani, A. *Pakistan Journal of Medical Sciences* **2017**, *33*, 1138. [Crossref]

[28] Kuder, J. F.; Richardson, M. W. *Psychometrika* **1937,** *2*, 151. [Crossref]

[29] Crombach, L. J. *Psychometrika* **1951**, *16*, 297. [Crossref]

[30] Cicchetti, D. V. *Psychological Assesment* **1994**, *6*, 284. [Crossref]

[31] Allen, K.; Reed-Rhoads, T.; Terry, R. A.; Murphy, T. J.; Stone, A. D. *J. Eng. Educ.* **2008**, *97*, 87. [Crossref]

[32] Brennan, R. L.; Lee, W. *CASMA-Technical note* **2006**, [Link]

[33] Kaitz, H. B. *Psychometrika* **1945**, *10*, 127. [Crossref]

[34] Student, *Biometrika* **1908**, *6*, 1. [Crossref]

[35] Brown, T. L.: LeMay, Jr., H. E.; Bursten, B.E.: Woodward, P.: Langford, S.: Sagatis, D.: George, A.; Chemistry: the central science, 13th ed. London, U.K.: Pearson Higher Education, 2013.

[36] Atkins, P.; De Paula J.; Elements of physical chemistry, 11th ed. Oxford, U.K.: Oxford University Press, 2013.