Virtual Tools for Developing SCLAs: An Example of Concentration Analysis

Andrew J. Mason
Department of Physics, Astronomy, and Engineering, University of Central Arkansas, Conway, AR 72035

Liam G. E. McDermott
Department of Physics, University of Connecticut – Avery Point, Groton, CT 06340

Mayuri Gilhooly and James Vesenka
School of Molecular and Physical Sciences, University of New England, Biddeford, ME 04005

Dawn Meredith
Department of Physics and Astronomy, University of New Hampshire, Durham, NH 03824

D. J. Wagner
Department of Physics, Grove City College, Grove City, PA 16127

Daniel Young
Department of Physics, Westminster University, Salt Lake City, UT 84105

Rebecca Lindell
Tiliadal Solutions, Lafayette, IN 47901

Abstract: Standardized Conceptual Learning Assessments (SCLAs) measure individuals' conceptual understanding using distractor-driven multiple-choice items. Concentration analysis (CA), a quantitative technique that evaluates the concentrations of each item's response distribution allowing researchers to determine how many models are triggered by each item [1], and can be utilized to help establish the validity of distractor driven items [2]. As part of the development of the Fluids Conceptual Evaluation (FCE), a distractor-driven two-tier multiple-choice evaluation instrument, the authors evaluated the concentration factors of a set of responses from the FCE's preliminary pilot test results. Results of this analysis show how many different mental models are triggered by each item's first tier. This paper demonstrates how the use of CA verifies recommended models, and suggests a similar model structure for concentration deviation. Recommendations for future use of concentration analysis will be made accordingly. Supported by NSF Award # 2021059 (AAPT), 2021261 (UNE) and 2021273 (UNH).

2025 PERC Proceedings edited by Pawl, Zwolak, and Leak; Peer-reviewed, doi.org/10.1119/perc.2025.pr.Mason

I. INTRODUCTION

When assessing the efficacy of a teaching method, class structure, or curriculum, conceptual learning assessments are a vital tool in an instructor's or researcher's repertoire. Traditionally, assessment development uses factor analysis, a parametric approach to analyze response patterns from data samples, in order to categorize a given assessment's items into factors. The development of distractor-driven conceptual assessments, however, needs a different approach, as factor analysis does not take into account how the context may differ between respondents based upon interpretations. This consideration of different items' respective context dependence also should be incorporated longer-term into the wider goal of creating standardized conceptual learning assessments (SCLAs).

One such way to assess the efficacy of an assessment is to use concentration analysis (CA) [1, 2]. CA is an analytical method for quantifying the quality of an assessment. It was developed as a tool to help design more effective distractor item response choices and probe common alternate conceptual models that students may have [3], which has also been a concern for existing instruments such as the FCI [4]. CA makes three key assumptions [1]: "1) Memory is associative; 2) Cognitive responses are productive; [and] 3) Cognitive responses are context dependent (including the context of the student's state of mind!)." These assumptions permit CA users to not settle for a unidimensional variable (whether students are either right or wrong about an item's topic), but instead allow for nuance about students' incorrect answers – in other words, to consider whether they are intelligently applying analytical frameworks to preconceived ideas, rather than just assume they do not understand said ideas and leave it at that. The goal of CA is therefore to provide additional information and context to student responses beyond what is covered by individual and collective test scores. CA achieves this goal by quantifying both how many students chose the correct answer and how concentrated student responses were on every possible answer. It also permits instrument developers of distractor-driven assessments to identify the respective strength of each item's distractors and make decisions about editing or omitting items accordingly.

Using CA to quantify how students respond to items in an assessment allows researchers to make qualitative decisions about which items are the most suitable to include in a finalized product. CA's main strength in this is to indicate whether a given item is too easy, whether a distractor choice for a given item is too strong, and whether a given item is simply difficult in terms of a seemingly random distribution of responses. By analyzing how students choose incorrect answers, educators can better address student alternative conceptions by knowing what phenomenological primitives [5], naïve misconceptions, etc. that students are using in their problem solving. Research shows [2] that CA is helpful to developers of a distractor-driven assessment.

The beta version of the Fluids Concept Evaluation (FCE) assesses pre-post conceptual understanding of several constructs concerning fluid mechanics, for the long-term purposes of assisting instructors of Introductory Physics for Life Science (IPLS) with development of fluids topics within their curricula. In this paper, we discuss CA as applied to FCE pilot data.

The FCE, in its current form, consists of 59 two-tier questions covering different aspects of fluid mechanics. The two-tier model [6] for each item consists of a first question (referred to hence as "Tier 1") which asks for a straightforward correct answer among a series of MC choices, followed by a second question (referred to hence as "Tier 2") which asks the respondent for the rationale behind choosing the preferred answer to the Tier 1 question, again from a list of possible rationales in MC format. Each question in either tier had different numbers of choices available, ranging from four choices in several cases to 11 choices for one question. In this paper, we will focus on Tier 1 questions for the items only.

As guided within our longer-term goal of establishing CA as a necessary consideration for developing SCLAs, we developed the following research questions. First, we are interested to see whether CA detects similar models of concentration factor and score as the models recommended in prior literature [1], and to what extent the FCE's first-tier items line up with the recommended models. Second, we wish to determine whether concentration deviation – a related measurement that specifically focuses upon incorrect responses only – can be analyzed with similar models as with concentration factor. Third, we seek to demonstrate an example of identifying FCE items that need to be addressed from the above analysis.

II. PROCEDURE

A. Data Collection

Data was collected as part of the FCE evaluation process. Participants were selected from physics departments of two universities in New England, USA (one research-focused, one teaching-focused) and one public research university in the South, USA. Institutions were selected based on knowledge of their introductory physics curriculum covering (at least as a subset of the full curriculum) fluid mechanics. A total of 290 participants responded to the survey.

As requiring students to respond to all 59 two-tier questions would require too much course time, participants were instead assigned a random set of 8 two-tier items to respond to. This approach ensured that each of the 59 items would receive between 35 and 45 responses. The data set was then compiled into a .CSV file and cleaned for analysis, de-identifying all data and condensing each item's respective responses.

B. Concentration Analysis

In addition to score S (the percent of students who answer a given question correctly), Ref. [1] developed a definition of the concentration factor, denoted as "C," given by Eqn. 1. Here, m is the number of choices for a given question, N is the total number of respondents, and n_i is the number of respondents who picked the i^{th} response choice for a given item. The gist of Eqn. 1 is a square root of the sum of squares over all n_i values, divided by the total number of responses for an item; other terms in Eqn. 1 collectively normalize the resulting C value from 0 to 1, with 0 being the theoretical limit of all choices having the same n_i value (such that $n_i = N/m$ for all i), and 1 being the limit of all students choosing a specific response choice. Ref. [2] gives guidelines of what C value characterizes a valid assessment item: a value of C between 0 and 0.20 is considered "low," a value between 0.20 and 0.50 is considered "medium," and a value above 0.50 is considered "high," such that a C score of 0 translates to evenly distributed responses across all choices and a C score of 1 indicates that all respondents picked the same choice. Ref. [1] combines this C factor analysis with a discrimination factor from Item Response Theory [4], which is not the focus of this paper.

$$C = \frac{\sqrt{m}}{\sqrt{m}-1}\left(\frac{\sqrt{\Sigma_{i=1}^{m} n_i^2}}{N} - \frac{1}{\sqrt{m}}\right) \quad (1)$$

The concentration deviation (denoted as "Γ"), given with Eqn. 2, is defined by removing the concentration of incorrect responses. The equation for defining Γ therefore is an adjustment of Eqn. 1 for defining C in that way. Here, $n_{correct}$ indicates the number of students who picked the correct response, who are removed from the sample of students; all other variables in Eqn. 2 have the same definitions as in Eqn. 1. The reason for also examining this measurement is that the output of C couples the correct response rate – in other words, the score (S) for a given item – with the incorrect response rates, which translates to upper (C_{max}) and lower (C_{min}) bounds for a given item's concentration rate (see Fig. 1, specifically the pattern of data). Γ is decoupled from S by effectively removing the number of students who chose the correct response from the rest of the student population. Thus, Γ is not dependent on S, allowing for an unbiased look at solely the incorrect response rates, and removing the upper and lower bounds for a given item's concentration deviation (see Fig. 2). Low, medium, and high values of concentration deviation are similarly defined [1, 2] as they are for C discussed above (low = 0 to 0.20; medium = 0.20 to 0.50; high = 0.50 to 1), with a similar meaning of the degree of focus on a particular incorrect choice (high value) versus random distribution across all incorrect choices (low value).

$$\Gamma = \frac{\sqrt{m}}{\sqrt{m}-1}\left(\frac{\sqrt{(\Sigma_{i=1}^{m} n_i^2) - n_{correct}^2}}{N - n_{correct}} - \frac{1}{\sqrt{m}}\right) \quad (2)$$

In order to categorize score and concentration factor, this paper will report the values of score first and concentration second, in terms of high (H), medium (M), or low (L). For example, an item with low score and medium concentration would be categorized as LM. Using recommended guidelines [1, 2], an item needs to have pre and post concentration factor of M or H to be considered valid. Dega and Govender [7] iterate on this S-C characterization to describe three different types of student modelling of a concept. HH and LH items show one dominant model, LM, MM, ML, MH items all show dual-dominant or dual-unpopular models, and LL items show no dominant model.

C. CA calculation

CA was programmed by Author 1 to calculate the values needed for Eqns. 1 and 2, organize the values into columns of data corresponding to the Tier 1 questions for each FCE item, and create S-C and S-Γ plots. More specifically, due to the preliminary nature of the FCE, departures from typical examples of concentration analysis in prior literature were expected and are being addressed. A prominent example, pertinent to this paper, is that the number of choices "*m"* and the number of student responses "*N*" both vary from question to question, unlike traditional assessments that require the same values for *m* and *N* respectively. An algorithm to count the number of unique responses to report as *m*, for example, can fail in cases which certain choices (most commonly, the last choice in the order presented to respondents) receive zero responses, and therefore are not counted, giving an erroneous value for *m*.

III. RESULTS

A. Concentration Factor Distribution

Table I shows the scores, concentration factors, and concentration discriminations for the Tier 1 questions in each of the 59 items of a preliminary draft of the FCE survey, as well as the response rates per question. For the scores, we follow the guidelines from Ref. [1] as a fraction (L: 0 – 0.4, M: 0.4 – 0.7, H: 0.7 – 1). As discussed above, a key detail is that, due to the FCE's current design, the value of *m* differs from question to question, as the questions have different numbers of choices, which will affect the calculations of C and Γ for each item. Note also that the number of students responding to each item (*N*) is not constant, varying between items from $N = 35$ to $N = 43$, as a side effect of the data collection methods discussed earlier. Table I also sorts the items according to the prescribed high, medium, and low values for S, C, and Γ, categorizing them by S-C model and subtype, as defined by Ref. 1, as well as consider an analogous S-Γ model and subtype along the same lines.

Most items fit in one of the prescribed S-C models, with seven items of the one-model type (four with the correct answer dominant, three with a dominant distractor), 28 items of the two-model type (19 items with the correct answer and one incorrect answer, nine items with two incorrect responses) and 17 items of the non-model type, with random

concentration. Of the seven items that did not fit a prescribed model, four (15, 28, 33, and 39) were in between the one-model subtype of a dominant correct answer and a two-model subtype of the correct answer and one particular incorrect answer. Items 28, 33, and 39 had three available incorrect choices, either one choice (33, 39) or two choices (28) of which dominated among incorrect responses, while item 15 had a sufficiently dominant distractor choice to allow for high C while keeping the score under the high threshold. Of the dominant correct answer subtype, two of the three items (1, 22) exhibited a low concentration deviation, in that the incorrect students were roughly evenly distributed across the incorrect choices, while the other item (14) had high concentration deviation, in that one of the three incorrect choices was picked by eight out of 10 students who were incorrect on this item. Four items (3, 30, 50, 52) were of the dominant incorrect answer subtype of one-model. These items all had only four choices, such that the dominant incorrect answer is a distractor that might suggest a "p-prim" [5], or some alternate form of misconception, is at play.

We also introduce in Table I an analogous set of S-Γ types and subtypes to the recommended S-C types and subtypes. In this preliminary stage, this classification is modeled to interpret the concentration of incorrect responses in relation to the percentage of correct responses. The degree to which the S-C type matches the S-Γ type – in other words, the degree to which the value for C is the same category as the value for Γ - is stronger for lower scores than for medium scores, and correspondingly stronger for medium scores than for higher scores. Of the 29 items with a low value for S, 26 had the same value category for C and Γ (90%). In comparison, this was the case for 15 out of 24 items with a medium S (62.5%) and for two out of the six items with a high S (33%). A high S means that relatively few respondents choose from all incorrect answers, meaning that C is more strongly dependent upon S, therefore G is more independent

TABLE I: Number of responses (*N*), number of item choices (*m*), score (S), concentration factor (C), concentration deviation (Γ), S-C type (S-C), and S-Γ type (S-Γ) for each of the 59 preliminary items for the FCE within the student sample. Based on recommended scales [1,2], "high" values are in bold, "medium" values are in normal font, and "low" values are in italics. The patterns for low (L), medium (M), and high (H) values are presented in the S-C row in the form of "S level, C level," and similarly for the S-Γ row in the form of "S level, Γ level."

Item	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
N	41	39	39	37	42	41	37	40	43	40	42	41	35	38	37
m	8	4	4	4	7	4	4	7	4	8	4	4	11	4	7
S	**0.88**	0.62	*0.18*	0.59	*0.31*	*0.39*	0.62	0.53	*0.40*	*0.15*	*0.17*	0.46	*0.09*	**0.74**	0.68
C	**0.82**	0.34	**0.58**	0.30	*0.38*	*0.16*	0.36	0.35	*0.16*	*0.08*	0.38	*0.18*	*0.12*	**0.54**	**0.51**
Γ	0.24	0.27	**0.85**	*0.16*	**0.61**	0.29	0.37	0.32	0.29	*0.10*	**0.54**	0.25	*0.14*	**0.59**	*0.13*
S-C	H,H	M,M	L,H	M,M	L,M	L,L	M,M	M,M	L,L	L,L	L,M	M,L	L,L	H,H	M,H
S-Γ	H,M	M,M	L,H	M,L	L,H	L,M	M,M	M,M	L,M	L,L	L,H	M,M	L,L	H,H	M,L
Item	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
N	38	38	39	39	40	39	42	41	40	37	39	42	39	40	42
m	4	4	4	8	5	8	4	5	4	7	4	4	4	5	4
S	*0.26*	*0.21*	0.56	*0.36*	0.50	*0.15*	**0.74**	*0.15*	*0.18*	0.41	0.49	0.40	**0.72**	*0.28*	*0.12*
C	*0.12*	0.21	0.29	0.32	0.24	*0.09*	**0.51**	*0.12*	*0.10*	0.20	*0.20*	*0.17*	0.48	0.26	**0.60**
Γ	0.21	0.34	0.32	0.47	*0.17*	*0.11*	*0.17*	*0.16*	*0.13*	*0.17*	0.25	0.32	*0.20*	0.44	**0.75**
S-C	L,L	L,M	M,M	L,M	M,M	L,L	H,H	L,L	L,L	M,M	M,L	M,L	H,M	L,M	L,H
S-Γ	L,M	L,M	M,M	L,M	M,L	L,L	H,L	L,L	L,L	M,L	M,M	M,M	H,L	L,M	L,H
Item	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**45**
N	38	38	38	40	36	41	42	43	41	38	39	40	43	35	45
m	4	9	4	6	6	8	7	5	4	4	4	6	9	7	5
S	0.55	*0.26*	**0.71**	*0.28*	*0.22*	*0.17*	0.55	0.53	**0.71**	0.45	0.51	*0.25*	*0.12*	*0.31*	*0.24*
C	0.29	*0.16*	0.48	*0.13*	0.22	*0.09*	0.34	0.34	0.47	0.20	0.25	0.26	*0.13*	*0.10*	*0.13*
Γ	0.38	*0.19*	0.37	*0.19*	0.33	*0.12*	*0.10*	0.47	*0.19*	0.35	0.36	0.40	*0.16*	*0.07*	0.22
S-C	M,M	L,L	H,M	L,L	L,M	L,L	M,M	M,M	H,M	M,M	M,M	L,M	L,L	L,L	L,L
S-Γ	M,M	L,L	H,M	L,L	L,M	L,L	M,L	M,M	H,L	M,M	M,M	L,M	L,L	L,L	L,M
Item	**46**	**47**	**48**	**49**	**50**	**51**	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	
N	34	34	44	36	37	38	40	40	38	36	40	39	39	41	
m	7	6	7	4	4	4	4	5	4	5	5	4	8	7	
S	*0.38*	0.47	*0.27*	0.58	*0.05*	*0.26*	*0.10*	0.58	0.63	0.64	0.58	0.54	*0.21*	*0.34*	
C	0.21	0.26	0.25	0.32	**0.69**	*0.08*	**0.76**	0.38	0.38	0.42	0.37	0.33	*0.12*	*0.15*	
Γ	0.22	0.23	0.37	0.40	**0.74**	*0.15*	**0.94**	0.49	0.41	0.28	0.42	**0.64**	*0.16*	*0.15*	
S-C	L,M	M,M	L,M	M,M	L,H	L,L	L,H	M,M	M,M	M,M	M,M	M,M	L,L	L,L	
S-Γ	L,M	M,M	L,M	M,M	L,H	L,L	L,H	M,M	M,M	M,M	M,M	M,H	L,L	L,L	

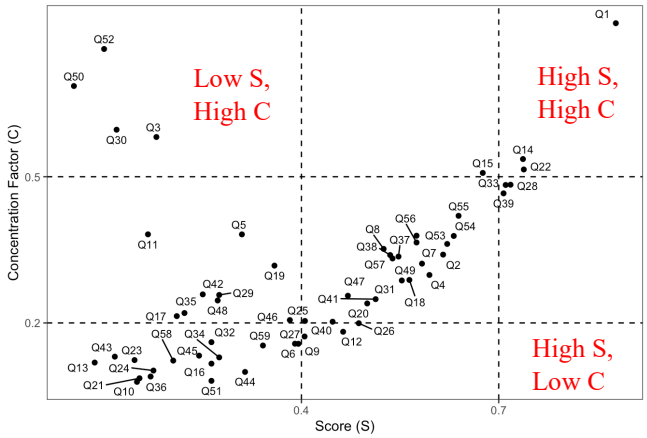

FIG. 1: Score-concentration factor (S-C) plot for items in Table I.

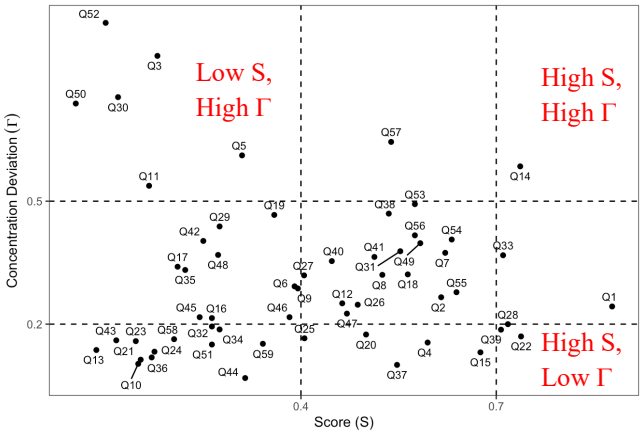

FIG. 2: Score-concentration deviation (S-Γ) plot for Table I items.

of C. A low S on the other hand, for which most respondents choose an incorrect response anyway, means that C is less dependent on S and more closely related to Γ.

B. S-C Plot and S-Γ Plot

In an S-C graph, C is plotted against S for each item in a scatter plot, with regions marked by S-C pattern (see Fig. 1). Note that the pattern of responses appears to be contained within an approximate upper bound and a lower bound. These bounds correspond to C_{max} and C_{min}, respectively, as explained above. The reason for these bounds is that a high concentration indicates that students are predominantly choosing a single response, whether the correct one (bounded at S = 1, C = 1) or an incorrect choice (bounded at S = 0, C = 1). In addition, it is impossible for evenly spread concentrations (low C) to have high S values. However, due to the calculation of C depending on *m* and *N* (see Table I), the bounds are not as well defined as they would be in cases where *m* and *N* are each the same across all items. Accordingly, the bounds are not explicitly shown in Fig. 1.

Similarly, Γ is plotted versus S in a S- Γ plot (see Fig. 2). As previously discussed, Fig. 2 does not have the bounds similar to those of the S-C plot in Fig. 1, as the calculation of Γ is dependent upon the distribution of incorrect responses without being also dependent upon score. As discussed earlier, the S-Γ pattern for lower scores overlaps strongly with the S-C pattern for lower scores, e.g. questions 3, 30, 50, and 52, of the dominant-incorrect subtype of one-model. A high concentration deviation in those cases reflects a dominant incorrect choice one-model subtype, given that adjusting to remove the score from this analysis has little change from C to Γ for these items.

IV. CONCLUSIONS AND DISCUSSION

Overall, CA as applied to the beta-version of the FCE's Tier 1 questions shows that the vast majority of items fit within one of the recommended S-C type models [1] and can be analyzed accordingly, as the few remaining items appear to be expressible as a "middle state" between two given models.

As seen in Fig. 2, a similar pattern of S-Γ type models exist; there appears to be more overlap of S-Γ models with S-C models – for example, in the sense that a medium value of C corresponds to a medium value of Γ – with lower scores, indicating that S-Γ type is more of a separate measure from S-C type with medium to high scores, such that it can be utilized more strongly as a different measurement at those score values. We have also been able to identify a few items that have a strong distractor (low S, high C, high Γ) that suggest a p-prim or some other misconception at play.

A couple future directions within this study are clear. The items featured here are only Tier 1 questions for the full preliminary FCE, in which each full item has two tiers of questions answered in succession, e.g. item 1 Tier 1 is asked and then item 1 Tier 2 immediately follows. Tier 2 item choices similarly have a correct explanation presented alongside a number of incorrect explanations in MC format. It remains to be seen whether CA is similarly useful for the Tier 2 questions; a concern is that the responses to Tier 2 items are conditional upon the Tier 1 items' respective responses. Each of the 59 items are sorted into one of 10 "constructs," for which a construct is defined as a subgroup of items that fit into a specific topic within fluids. Concentration analysis presented here will assist comparative analysis of items within each construct, guiding qualitative analysis on whether item design, student response patterns, etc. contributes to the differences between items, with the goal of selecting and finalizing a normalized set of items for future development.

V. ACKNOWLEDGEMENTS

The authors would like to thank the rest of the Fluid RIPLS team for valuable feedback on this paper. In addition, we wish to extend our sincere gratitude to the three institutions who volunteered to collect preliminary FCE pilot tests.

[1] Bao, L., & Redish, E. F. (2001). Concentration analysis: A quantitative assessment of student states. *American Journal of Physics*, *69*(S1), S45-S53.
[2] Lindell, R. S., & Olsen, J. P. (2002). Developing the Lunar Phases Concept Inventory. In *2002 PERC Proceedings* [Boise, ID, August 7-8, 2002], edited by S. Franklin, K. Cummings, and J. Marx, doi:10.1119/perc.2002.pr.011. <https://www.compadre.org/Repository/document/ServeFile.cfm?ID=4323&DocID=1162>.
[3] Bao, L., & Redish, E. F. (2006). Model analysis: Representing and assessing the dynamics of student learning. *Physical Revew Special Topics: Physics Education Research* 2(1), 010103.
[4] Wang, J., & Bao, L. (2010). Analyzing force concept inventory with item response theory. *American Journal of Physics*, *78*(10), 1064-1070.
[5] diSessa, A. A. (1993). Toward an Epistemology of Physics. *Cognition and Instruction*, *10*(2–3), 105–225.
[6] e.g. Sıbıç, O., Akçay, B., & Arık, M. (2020). Review of Two-tier Tests in the Studies: Creating a New Pathway for Development of Two-tier Tests. *International Journal of Contemporary Educational Research,* 7(2), 81-98.
[7] Dega, B. G., & Govender, N. (2016). Assessment of students' scientific and alternative conceptions of energy and momentum using concentration analysis. *African Journal of Research in Mathematics, Science and Technology Education*, 20(3), 201-213.