**RESPONSE TO REVIEWERS**

**REVIEWER B:**

1. The authors suggested solving of the problem by using chemical simulations and conceptual change text. Although six-simulation are mentioned in the manuscript, there is no closer explanation what simulations include. Please, explain them in more details. The same is valid for the conceptual change text.

Detailed explanation related to developing process and properties of the simulations (p.7) and conceptual change texts (p.7) was made in the manuscript.

2. Since I still believe in chemistry as experimental science, please modify conclusion in sense that simulations are more than welcome in teaching chemistry but along with experiments performed in the lab.

The phrase that laboratory is vital and indivisible part of the chemistry teaching was added into the conclusion section of the paper (p.20).

**REVIEWER C:**

1. The prior knowledge on chemical equilibrium of the study sample (before conducting pre-test) is not described. Since pre-test and post-test were identical, and the intervention was carried out between two testing, I wonder whether the results of pre-test indicate the existence of alternative conceptions, as suggested by the authors, or incorrect answers are rather the result of the lack of knowledge. Is it to be expected that students give correct answers, or opposite, to possess alternative conceptions if the pre-test was conducted before studying the contents of chemical equilibrium? Inaccurate answers need to be distinguished from alternative conceptions. Providing information about students’ prior knowledge regarding chemical equilibrium could help in explaining this ambiguity.

Detailed information about the prior knowledge level of the students was added in the headings of “procedure”. (p. 9).

1. I find scoring system a little bit puzzling. To score 2 points both tiers must be correctly chosen; to score 1 point only one tier should be correctly chosen and to score 0 points both tiers must be incorrectly chosen (as if the test was common multiple choice). However, in the Results section the authors provided only mean values, which cannot be indicator of understanding. The reader gains no insights about percentages of correct combinations of answers (both-tier scores) which are crucial when discussing about students’ understanding. Since the obtained mean values are quite low, even after the intervention (40.2% for post-test, 38.0% for delayed post-test) I am not convinced that such values are the result of correct combinations of answers.  Given that one of the core goals of this study was to determine the effect of CS-CCT instruction on students’ understanding, achieved scores in EG should be mainly the result of paired scores in order to conclude that CS-CCT had positive impact on students’ understanding. I'd recommend authors to include and discuss these results in more detail.

The paired scores of the EG and CG students were added as Table IV (p.11-12).

1. The authors have used two-tier test which proved to be more reliable and valid than common multiple choice tests, according to a large body of literature in the field of evaluation and assessment. However, instrument used in this research is characterized by only one statistical parameter (Cronbach alpha coefficient). Item difficulty and item discrimination indices are commonly reported in similar research studies, and should be included as well.

Information on characteristics of the CECT was added as Table I (p. 5).

1. What is somewhat surprising is pretty modest paragraph on the results of delayed test, as these results, according to my opinion, indicate the most important contribution of this study. Namely, although statistically significant, difference between EG and CG scores on the post-test is not high (40.2% and 35.1%, respectively). On the other hand, this difference  
   increases on delayed test and it is significant as it may indicate the difference in the quality of the acquired knowledge among groups. Along with both-tier scores, this is the part of the paper that should provide most evidence about the effectiveness of applied instruction.

The statistically calculations were made again based on the students scores. Unfortunately, we used false statistical data in the first version of the manuscript. Based on the reviewer C cautions, we made t test and ANCOVA again and saw that the results were different from the first ones. So, we added the re-calculated data on the paper as Table II and Table III (p. 10-11).

1. The manuscript ends without any direct implications for either future  
   research or teaching. Limitations are not described as well. This part of  
   the paper needs to be enriched.

Implications and limitations were added on tha manuscript (p. 20-21).

1. The authors introduce a system of labeling: CS for Simulations, CCT for Conceptual Change texts, EG for Experimental Group and CG for Control Group (occasionally referred to as comparison group), but in the subsequent discussion, these labels are used sporadically and inconsistently (e.g.  
   lines [229, 232, 238-240, 242 etc.]). The system should be used each and every time, or removed from the manuscript.

All of the labels were corrected on the whole text.

1. Line [126]: “Distracters for the questions were taken from the related literature” lacks relevant literature citations.

Required literature and information on CECT’s conceptual areas were added (p.5).

1. In provided examples of questions and CCT, forward arrows should be  
   replaced by equilibrium arrows.

It was made (p.5).

9. Formulation of alternative conception (Table IV): “The rate of forward reaction is greater than the reverse reaction one” should be more precise. I suggest one of the following formulations: (i) The rate of forward reaction is greater than the reverse reaction one at equilibrium; (ii) The rate of forward reaction is always greater than the reverse reaction one.

It was made by adding the word of “always” (p. 14)