# WIP: Psychometric Analysis of Electric Circuit Concepts Diagnostic (ECCD) Basic Test Items: A Classical Test Theory Approach

[1]Tamlyn Lahoud; [2]Olanrewaju P. Olaogun; [1]Shiyu Wang; [3]Nathaniel Hunsu; [3]Adel Al-Weshah; [3]Kun Yao

[1]College of Education
University of Georgia, Athens, United States

[2]School of Engineering and Computational Sciences
Merrimack College, North Andover, United States

[3]College of Engineering
University of Georgia, Athens, United States

***Abstract*— Concept inventories are diagnostic instruments designed to assess student understanding and identify common misconceptions in a specific domain or discipline. They consist of multiple-choice questions or tasks that target key concepts and principles in the subject matter. The ECCD was developed to diagnose student misconceptions of electric concepts like power, energy, and potential difference. For concept inventories to be effective in measuring misconceptions, the items on the test must possess two essential qualities: validity and reliability. We administered the ECCD to 270 undergraduate engineering students in the summer and fall of 2023 and the test items are examined using the classical test theory framework. Results from the analysis show that most of the items on the test perform well in terms of item difficulty, item discrimination, and reliability. Findings from this work-in-progress analysis pointed to a few items that might need revision or be discarded on the final test. The implication of these findings is also reported.**

***Keywords—Psychometric Properties, Electric Circuits, Concept Inventory***

## I. INTRODUCTION

The field of electrical engineering is replete with abstract, complex, and counter-intuitive concepts that students have to learn to achieve competency as electrical engineers [1]. Researchers develop and validate knowledge diagnostics or assessment tools such as the Electric Circuit Concepts Diagnostic (ECCD) to assess misconceptions in students' prior knowledge and knowledge gain in science domains. The ECCD was designed to assess students' misconceptions and test their understanding of fundamental electrical circuit concepts [2, 3]. While concept inventories have been developed to detect and identify various misconceptions that students hold in different domains of engineering sciences and education, interpreting the scores from these inventories can be challenging. The ambiguity arises from the difficulty in distinguishing between a lack of knowledge and the presence of misconceptions based solely on students' responses to the inventory items. This limitation complicates the effective use of concept inventories in educational assessments. The Electric Circuit Concept Diagnostic (ECCD) test addresses this issue through a three-tier assessment approach, which evaluates not only students' content knowledge but also their reasoning behind the chosen answers and their confidence level in those answers [3]. Furthermore, knowledge diagnostics not only help in identifying misconceptions that students have but also serves as a feedback mechanism for educators to adjust instructional strategies. Such assessments are integral to ensuring that all impediments to learning caused by students' faulty prior knowledge are removed and that students can conceptually grasp the lessons being taught.

Psychometric analysis is a cornerstone of educational assessment, providing essential insights into test validity, reliability, and the effectiveness of educational tools [4],[5]. Validity refers to the extent to which an inventory accurately measures the intended concept or construct [6]. This involves ensuring that the items on the test are directly related to the target concepts and are not influenced by irrelevant factors. Reliability, on the other hand, refers to the consistency of results produced by a measurement tool when administered multiple times. Reliability of measures is crucial in educational research where consistent and accurate measurement is essential [7].

The ECCD is currently being developed by our research team to assess students' misconceptions about electric concepts. This work-in-progress study documents a preliminary study that aims to evaluate the psychometric properties of ECCD test through a detailed analysis of the difficulty, discrimination, and reliability of items on the test. By employing robust psychometric methods, this article seeks to illuminate how well the test's items measure what they intend to and how their results can inform educational practice. Through this examination, we intend to provide a comprehensive overview of the psychometric foundations of the ECCD, thereby aiding educators, curriculum developers, and researchers in helping students to overcome their misconceptions and enhancing the educational experiences of electrical engineering students.

## II. THEORETICAL FRAMEWORK

Classical Test Theory (CTT) is built upon several foundational principles and assumptions that focus on the measurement of latent traits or abilities through observed test scores. The CCT posits that each test score is composed of a true score (the actual measure of the trait or ability) and an error score (random and unpredictable variations that affect the test score). The true score reflects the stable component of an individual's performance, whereas the error represents the measurement noise [8]. One central concept in CTT is reliability, defined as the proportion of variance in the observed scores that can be attributed to the true scores. Reliability is often quantified through various methods, including the parallel-test reliability formula, which considers test length and a parameter of the underlying population distribution of abilities [8].

The CTT operates under the assumption that the relationship between observed scores and true scores is linear. Similarly, the errors associated with measurement must be independent of each other and of the true score. And finally, the error variances must be constant across all test items [9, 10]. Under the CCT, factor analysis is often used to develop scales and mention the dimensionality of the data which provides insight into the validity of the scale.

Despite its simplicity, CTT forms the basis for many of the standard practices in the development and evaluation of tests and is essential for understanding measurement properties like validity and reliability [11].

## III. METHOD

### A. Participants

This study's participants were undergraduate students enrolled in different engineering programs at a large public research university in the US. We administered the ECCD to 270 undergraduate engineering students who gave their consent to participate in the summer and fall of 2023. Seven responses were missing and only the 263 participants who responded to all the items were included in this analysis.

This ECCD includes 30 multiple-choice test items that assess 10 electric circuit concepts. The test was administered as three equivalent sub-scales of the tests. Each sub scale contains 10 questions that measure each of the concepts assessed on the test. These sub-scales were administered to three different groups of students (See Table I). The data explored in this report include the original responses (options) selected by the students for each question and their binary (scored) responses, indicating whether their answers were correct or incorrect.

Using the Classical Test Theory (CTT), we conducted item analysis, which included item calibration (difficulty and discrimination estimation), distractor analysis, and reliability testing.

TABLE II: ITEM DIFFICULTY ESTIMATES AND 95% CONFIDENCE INTERVALS

| Item | Mini-test 1 | | Mini-test 2 | | Mini-test 3 | |
|---|---|---|---|---|---|---|
| | *Difficulty* | *95% CI* | *Difficulty* | *95% CI* | *Difficulty* | *95% CI* |
| Question.1 | 0.545 | 0.467;0.622 | 0.400 | 0.307;0.493 | 0.699 | 0.616;0.782 |
| Question.2 | 0.750 | 0.683;0.817 | 0.600 | 0.507;0.693 | 0.819 | 0.750;0.889 |
| Question.3 | 0.795 | 0.732;0.857 | 0.400 | 0.307;0.493 | 0.458 | 0.368;0.548 |
| Question.4 | 0.616 | 0.540;0.692 | 0.333 | 0.244;0.423 | 0.482 | 0.392;0.572 |
| Question.5 | 0.357 | 0.283;0.432 | 0.573 | 0.479;0.667 | 0.398 | 0.309;0.486 |
| Question.6 | 0.375 | 0.300;0.450 | 0.653 | 0.563;0.744 | 0.518 | 0.428;0.608 |
| Question.7 | 0.625 | 0.550;0.700 | 0.480 | 0.385;0.575 | 0.482 | 0.392;0.572 |
| Question.8 | 0.804 | 0.742;0.865 | 0.573 | 0.479;0.667 | 0.554 | 0.464;0.644 |
| Question.9 | 0.875 | 0.824;0.926 | 0.573 | 0.479;0.667 | 0.602 | 0.514;0.691 |
| Question.10 | 0.545 | 0.467;0.622 | 0.347 | 0.256;0.437 | 0.639 | 0.552;0.725 |

## IV. DATA ANALYSIS

### A. Item Difficulty

For dichotomously scored items, under the CTT framework, item difficulty is defined as the proportion of examinees to answer that item correctly [12]

$$P^* = n_c/N$$

where $n_c$ is the number of correct responses and N refers to the total number of respondents for that item. Naturally this value ranges from zero to one, where higher values indicate easier items.

In addition, the variance can be calculated as

$S^2_{P^*} = P^*(Q^*)$ where $Q^* = 1 - P^*$.

From here, we can calculate the standard error as follows

$SE_{P^*} = \sqrt{S_{P^*}^2/N}$.

The standard error may be used to calculate the confidence interval of difficulty such that $P^* \pm z(SE_{P^*})$ where z is the standard normal distribution corresponding to the desired confidence level.

Difficult items have a difficulty level of 0–0.3; 0.3–0.8 is classified as a desirable item, and items with difficulty values over 0.8 are considered easy [13].

### B. *Distractor Analysis*

Sometimes, it is also helpful to consider how incorrect answers, known as distractors, are selected by examinees and identify distractors that perform poorly [12].

Finch and French [12] suggest three general procedures for conducting a distractor analysis.

1. First, consider inordinately attractive incorrect answers that examinees in the upper third of scores are more likely to choose as these could be considered misleading or poorly performing.

2. Then, insufficiently attractive incorrect response options, which are rarely selected (even by lower-scoring examinees) should be identified.

3. Lastly, items that are answered correctly more by low scoring than high scoring examinees should be reviewed for being misleading or incorrectly keyed.

### C. *Reliability Test*

There are various methods available for measuring the reliability of a test under a CTT framework [14]. The test/retest reliability is calculated based on the correlation of scores from the same examinees presented with the same test on two different occasions. The parallel-forms reliability is derived similarly to test/retest reliability, except two parallel tests are used instead of the same one. The third and most used type is internal-consistency reliability, such as Spearman-Brown reliability. We chose to use Spearman-Brown split-test reliability to evaluate each of the three mini-tests.

## V. Results and Observations

The items in mini-test 1 were answered correctly by 35.7%-87.5% of the examinees. Two items (Q.8, Q.9) were considered easy, eight were desirable, and none were difficult. All items had discrimination parameters above zero, which shows that the probability of answering an item correctly is positively related to the examinee's ability, so the items favored the higher-scoring group. Higher discrimination values are associated with better item quality, as it indicates that the item discriminates well between the higher and lower performers on the sub-scales of the test. From mini-test 1, two items (Q.8, Q.9) were poor quality, two (Q.2, Q.3) were fair and one (Q.6) was good quality. The remaining five items had an index of discrimination above 0.4, indicating that they are very good quality as seen in Table II.

The proportion of examinees who correctly answered each of the mini-test 2 items ranged from 33.3% to 67.3%. Therefore, all items fell within the desirable range for difficulty. Q.3 had a negative discrimination value, indicating that this item could be unclear or indexed incorrectly as it favors the lower scoring group. One item (Q.2) is considered good quality, and the remaining eight items are classified as very good quality items.

39.8%- 81.9% of the examinees answered the items in mini-test 3 correctly. Nine items were of desirable difficulty, one (Q.2) was easy, and none were considered difficult. All items had discrimination indices above zero. One (Q.2) was of good quality, and the remaining nine were considered very good quality(See Table III).

For the distractor analysis, the key provides the four response options from (a) to (d), and an asterix (*) indicates the correct response option.

The number and associated proportion of respondents who select each response option are presented in columns 'n' and 'Prop' respectively. The point-biserial correlation between each response and the total score with that item removed is also provided as a measure of reliability of the answer choices.

TABLE III: Item Discrimination Estimates

| Item | Mini-test 1 | Mini-test 2 | Mini-test 3 |
|---|---|---|---|
| Question.1 | 0.549 | 0.583 | 0.720 |
| Question.2 | 0.220 | 0.368 | 0.360 |
| Question.3 | 0.275 | -0.076 | 0.411 |
| Question.4 | 0.420 | 0.688 | 0.402 |
| Question.5 | 0.456 | 0.535 | 0.467 |
| Question.6 | 0.322 | 0.611 | 0.840 |
| Question.7 | 0.495 | 0.465 | 0.689 |
| Question.8 | 0.132 | 0.646 | 0.544 |
| Question.9 | 0.095 | 0.542 | 0.664 |
| Question.10 | 0.490 | 0.694 | 0.609 |

The discrimination value, 'discrim', is the upper proportion minus the lower proportion. It is expected that correct response options would have positive discrimination and incorrect options would have negative discrimination. To align with the discrimination estimation procedure used in this report, the upper and lower groups were selected as the upper and lower 20% of scores. These proportions are presented in 'Upper 20%' and 'Lower 20%' respectively.

The remaining 60% of examinees were split evenly, such that the 20-50% group proportion is in 'mid50' and the 50-80% in 'mid80'.

Nine of the items on the first sub-scale (Test 1) (See Table IV), had the expected response patterns. However, for Q.6, option (d) is a distractor since more examinees selected this response option than the correct one (c). It should be noted that within the highest-scoring group, more examinees (70%) selected the correct response, which indicates that it has been correctly keyed.

For mini-test 2, we found that there were two items (Q.3, and Q.4) for which more students selected an incorrect response option than the correct option. Overall, respondents were more than twice as likely to select the incorrect option (d) than the

TABLE IV: TEST 1 DISTRACTOR ANALYSIS

| Item | Key | n | Prop | Point-biserial | Discrim | Lower 20% | Mid 50% | Mid 80% | Upper 20% |
|---|---|---|---|---|---|---|---|---|---|
| Question.1 | (a) * | 61 | 0.545 | 0.234 | 0.744 | 0.256 | 0.476 | 0.738 | 1 |
| | (b) | 5 | 0.045 | -0.243 | -0.051 | 0.051 | 0.095 | 0.024 | 0 |
| | (c) | 4 | 0.036 | -0.316 | -0.103 | 0.103 | 0.000 | 0.000 | 0 |
| | (d) | 42 | 0.375 | -0.563 | -0.590 | 0.590 | 0.429 | 0.238 | 0 |
| Question.2 | (a) * | 84 | 0.750 | 0.136 | 0.436 | 0.564 | 0.762 | 0.857 | 1 |
| | (b) | 21 | 0.188 | -0.464 | -0.333 | 0.333 | 0.143 | 0.119 | 0 |
| | (c) | 5 | 0.045 | -0.289 | -0.077 | 0.077 | 0.048 | 0.024 | 0 |
| | (d) | 2 | 0.018 | -0.170 | -0.026 | 0.026 | 0.048 | 0.000 | 0 |
| Question.3 | (a) | 13 | 0.116 | -0.430 | -0.205 | 0.205 | 0.095 | 0.071 | 0 |
| | (b)* | 89 | 0.795 | 0.276 | 0.436 | 0.564 | 0.857 | 0.929 | 1 |
| | (c) | 5 | 0.045 | -0.357 | -0.128 | 0.128 | 0.000 | 0.000 | 0 |
| | (d) | 5 | 0.045 | -0.357 | -0.103 | 0.103 | 0.048 | 0.000 | 0 |
| Question.4 | (a) | 31 | 0.277 | -0.537 | -0.436 | 0.436 | 0.333 | 0.167 | 0 |
| | (b) | 4 | 0.036 | -0.134 | -0.051 | 0.051 | 0.048 | 0.024 | 0 |
| | (c) | 8 | 0.071 | -0.314 | -0.128 | 0.128 | 0.095 | 0.024 | 0 |
| | (d)* | 69 | 0.616 | 0.166 | 0.615 | 0.385 | 0.524 | 0.786 | 1 |
| Question.5 | (a)* | 40 | 0.357 | 0.054 | 0.444 | 0.256 | 0.238 | 0.429 | 0.7 |
| | (b) | 33 | 0.295 | -0.387 | -0.159 | 0.359 | 0.333 | 0.238 | 0.2 |
| | (c) | 20 | 0.179 | -0.326 | -0.205 | 0.205 | 0.238 | 0.167 | 0.0 |
| | (d) | 19 | 0.170 | -0.285 | -0.079 | 0.179 | 0.190 | 0.167 | 0.1 |
| Question.6 | (a) | 14 | 0.125 | -0.457 | -0.231 | 0.231 | 0.095 | 0.071 | 0.0 |
| | (b) | 3 | 0.027 | -0.297 | -0.051 | 0.051 | 0.000 | 0.024 | 0.0 |
| | (c)* | 42 | 0.375 | 0.074 | 0.521 | 0.179 | 0.429 | 0.452 | 0.7 |
| | (d) | 53 | 0.473 | -0.324 | -0.238 | 0.538 | 0.476 | 0.452 | 0.3 |
| Question.7 | (a)* | 70 | 0.625 | 0.190 | 0.515 | 0.385 | 0.476 | 0.857 | 0.9 |
| | (b) | 7 | 0.062 | -0.273 | -0.103 | 0.103 | 0.095 | 0.024 | 0.0 |
| | (c) | 27 | 0.241 | -0.504 | -0.310 | 0.410 | 0.286 | 0.095 | 0.1 |
| | (d) | 8 | 0.071 | -0.314 | -0.103 | 0.103 | 0.143 | 0.024 | 0.0 |
| Question.8 | (a)* | 90 | 0.804 | 0.101 | 0.285 | 0.615 | 1 | 0.857 | 0.9 |
| | (b) | 10 | 0.089 | -0.337 | -0.079 | 0.179 | 0 | 0.048 | 0.1 |
| | (c) | 11 | 0.098 | -0.340 | -0.179 | 0.179 | 0 | 0.095 | 0.0 |
| | (d) | 1 | 0.009 | -0.226 | -0.026 | 0.026 | 0 | 0.000 | 0.0 |
| Question.9 | (a) | 2 | 0.018 | -0.353 | -0.051 | 0.051 | 0.000 | 0.000 | 0 |
| | (b)* | 98 | 0.875 | 0.074 | 0.231 | 0.769 | 0.905 | 0.929 | 1 |
| | (c) | 1 | 0.009 | -0.121 | -0.026 | 0.026 | 0.000 | 0.000 | 0 |
| | (d) | 11 | 0.098 | -0.294 | -0.154 | 0.154 | 0.095 | 0.071 | 0 |
| Question.10 | (a) | 20 | 0.179 | -0.473 | -0.282 | 0.282 | 0.381 | 0.024 | 0 |
| | (b) | 17 | 0.152 | -0.398 | -0.256 | 0.256 | 0.143 | 0.095 | 0 |
| | (c) | 14 | 0.125 | -0.282 | -0.128 | 0.128 | 0.143 | 0.143 | 0 |
| | (d)* | 61 | 0.545 | 0.198 | 0.667 | 0.333 | 0.333 | 0.738 | 1 |

groups which could indicate that the item was incorrectly keyed, or the question and/or response options were misleading. Q.3 was already flagged as having negative discrimination, and the distractor analysis confirms that this item should be reviewed. Q.4 also had more respondents selecting the incorrect option (a), than the correct option (b); however, in the top two performance groups, more examinees selected the correct response. Interestingly, no respondents selected option (c). Since this item had a desirable difficulty and was considered very good quality, Q.4 may not require much revision. However, option (c) should be adjusted to increase its attractiveness.

Q.3 of Test 3 had slightly more respondents selected an incorrect response, however, in the top 50% of performers, more respondents selected the correct response.

Lastly, for the reliability test, mini-test 3 was the most reliable according to the Spearman-Brown split-test reliability as shown in Table V. This result aligns with the item analysis results which indicated that its items were of a desirable difficulty level and good quality, and the distractor analysis results, which showed only one item potentially requiring review. It is likely that if the recommended item revision is conducted, the reliability of Test 1 and 2 will also improve.

TABLE V: SPEARMAN-BROWN SPLIT-TEST RELIABILITY

| Test 1 | Test 2 | Test 3 |
|---|---|---|
| 0.429 | 0.315 | 0.628 |

## VI. Conclusion and Future Directions

This study examines the psychometric properties of the ECCD instrument developed to diagnose students' misconceptions about electric circuits concepts. Findings from this study suggest that there is a need for the revision of some items and removal of others. We hope that by the end of the development of this instrument, we will have produced an instrument capable of providing instructors with the requisite knowledge of student's misconceptions in electric circuits and be able to help them overcome them. This is a part of an ongoing project, and the ultimate goal is to host this inventory online and make it accessible to all instructors across the US and all over the world.

## Acknowledgment

The authors gratefully acknowledge the National Science Foundation for supporting this work under the Phase I (NSF IUSE #2021468). Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science

## References

[1] O. P. Olaogun, A. Skelton, M. Zafar, N. J. Hunsu, and I. A. Idowu, "A Systematic Review of Student Misconceptions about Electricity and Electric Circuit Concepts," *in 2023 IEEE Frontiers in Education Conference (FIE)*, 2023: IEEE, pp. 1-5.

[2] O. P. Olaogun, J. M. Foster, Z. Lin, A. Al Weshah, K. Yao, and N. J. Hunsu, "Determination of Students' Misconceptions using the Electric Circuit Concept Diagnostic (ECCD) Instrument," *in 2023 IEEE Frontiers in Education Conference (FIE)*, 2023: IEEE, pp. 1-5.

[3] N. Hunsu, K. Yao, A. Al-Weshah, O. Olaogun, "Work in Progress: The Electric Circuit Concepts Diagnostic (ECCD)," *Annual Proceeding of the American Society for Engineering Education, 2022.*

[4] J. Pérez, C. Vizcarro, J. García, A. Bermúdez, and R. Cobos, "Development of procedures to assess problem-solving competence in computing engineering," *IEEE Transactions on Education*, vol. 60, no. 1, pp. 22-28, 2016.

[5] E. W. Gordon, "Toward assessment in the service of learning," *Educational Measurement: Issues and Practice*, vol. 39, no. 3, pp. 72-78, 2020.

[6] E. Kaya, H. Deniz, and E. Yesilyurt, "Toward developing a valid and reliable assessment of adults' nature of engineering views," *J Eng Educ*, vol. 112, no. 3, pp. 634-673, 2023.

[7] D. Maric, G. A. Fore, S. C. Nyarko, and P. Varma-Nelson, "Measurement in STEM education research: A systematic literature review of trends in the psychometric evidence of scales," *International Journal of STEM Education*, vol. 10, no. 1, p. 39, 2023.

[8] C. Foster, "A generalized framework for classical test theory," *Journal of Mathematical Psychology*, vol. 96, p. 102330, 2020.

[9] N. Jorion, B. Self, K. James, L. Schroeder, L. DiBello, and J. Pellegrino, "Classical test theory analysis of the dynamics concept inventory," in *Proc. of the 2013 American Society of Engineering Education, Annual Conference*, 2013.

[10] R. L. Brennan, "Generalizability theory and classical test theory," *Applied measurement in education*, vol. 24, no. 1, pp. 1-21, 2010.

[11] M. Wu, H. P. Tam, T.-H. Jen, M. Wu, H. P. Tam, and T.-H. Jen, "Classical test theory," *Educational Measurement for Applied Researchers: Theory into Practice*, pp. 73-90, 2016.

[12] W. H. Finch and B. F. French, *Educational and psychological measurement*. Routledge, 2018.

[13] Y.-m. Chae, S. G. Park, and I. Park, "The relationship between classical item characteristics and item response time on computer-based testing," *Korean journal of medical education*, vol. 31, no. 1, p. 1, 2019.

[14] M. J. Allen and W. M. Yen, *Introduction to measurement theory*. Waveland Press, 2001.