International Society of the Learning Sciences

# Empirically-driven Multiple Choice Response Options

Aubrey Condor, University of California, Berkeley, aubrey_condor@berkeley.edu

**Abstract:** We introduce an empirical method for generating multiple choice (MC) response options for educational assessment. Developing quality response options for MC questions is labor intensive, especially when parallel items are necessary for repeated assessments or multiple test forms. We propose a data-driven, iterative approach for creating an item bank of MC questions by first collecting a sample of open-ended student responses that are subsequently rated by subject-matter experts. Contextual sentence embeddings of the observed student responses as well as the expert ratings are used in conjunction with hierarchical clustering to create MC options that span the desired ground truth rating levels specified by the item creator. We evaluate the quality of the automatically generated response options qualitatively, based on existing research relating to best practices for item writing.

## Introduction

Although the grading of multiple choice (MC) items is less time consuming than open-ended (OE) items, constructing quality response options for MC items can be more labor-intensive and difficult (Gierl, 2012). Previous research on item design suggests that response options should include feasible student conceptions about the topic of interest, whether they be correct or incorrect (Haladyna, 2010; Zimmaro 2010). In addition, related research suggests that MC options should be homogeneous in content but different in their level of correctness (Burton, 1990). Finding plausible distractors that reflect common student misconceptions and are contextually related to other options requires considerable effort and can be hindered by expert blindness (Nathan, 2003). We suggest that the application of artificial intelligence (AI) can help mediate such issues and aid in constructing items that exemplify these item writing guidelines.

The motivation for this work is to reduce a common time burden for educators, as well as suggest a more objective, data-driven method for creating quality MC response options for both single item creation, and the creation of a bank of parallel items. We propose an iterative item development method as such: first, desired items are created as OE items and a preliminary data collection is performed to obtain a sample of student responses. Once subject-matter experts rate the observed responses, we create vector embeddings of the responses, and use hierarchical clustering to automatically identify contextually similar response options that reflect real students' conceptions.

## Multiple Choice Response Option Generation

The data we used for this project was sourced from a 2019 field test of a critical reasoning in mathematics assessment developed by a team of researchers at the Berkeley Evaluation and Assessment Research (BEAR) center. The data included responses to 30 different mathematics related items. However, for this study, we focus on the 6 items with over 300 responses each. Student responses were each rated from 0 (fully incorrect) to 3 (fully correct) by multiple researchers at the BEAR center.

Our method for automatically generating MC response options is as follows. We first use Sentence-BERT (SBERT; Reimers, 2019) to create sentence embeddings of the collected responses. SBERT is a modified version of BERT that utilizes siamese and triplet network structures to create semantically meaningful sentence embeddings. Next, we reduce the dimension of the often-sparse sentence embeddings from vectors of size 768 to size 50 using truncated singular value decomposition (SVD). With the reduced sentence embeddings, we perform hierarchical clustering with the Ward linkage criteria. Hierarchical clustering consists of a family of algorithms that form clusters by merging or splitting data points in succession, and the Ward linkage method performs merging by minimizing the sum of squares within the clusters. We use the silhouette coefficient to search for the best number of clusters from 3 to 6.

Finally, we filter for clusters that include at least 20 responses to make sure there are enough responses from which to sample. We randomly sample response options from the user specified rating level categories to create an item version. If it is the case that the cluster from which we are sampling does not contain all rating categories needed, we sample from a different cluster. Once the final item responses are created, we recommend that a human intervenes to perform a quick pass of final necessary modifications to the responses such as making sure that all responses either end in a period or do not, removing clear grammar mistakes, etc. Item writing recommendations specify that answers should use similar language, there should not be any grammar issues or misspelled words to prevent obvious incorrect answer elimination (Burton, 1990).

 © ISLS

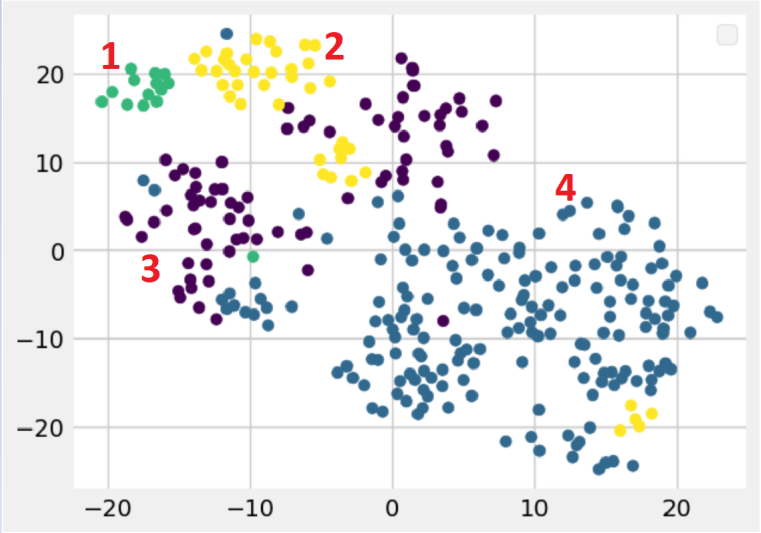
## Results and Discussion

As this project is a work-in-progress, we evaluate the goodness of the generated items qualitatively. Figure 1 shows a 2D visualization of the response clusters for an item that references a graph of an exponential function. For each of the four clusters, example responses are included as well as their respective ratings. The examples shown for clusters 2, 3, and 4 may serve as fixed response options for three different parallel items. The options span the 4 ratings levels, but the user can also specify to sample one fully correct answer (rated as 4) and the rest fully incorrect answers (rated as 0). Cluster 1 seems to contain "I don't know" type answers which should be avoided, and only answers rated as 0 or 2. Our method would not sample from this cluster because responses do not span the desired rating levels. Noticeable issues in some of the sampled response bundles include drastically uneven response lengths and the lowest rated responses being nonsense answers such as "I just guessed". However, as previously suggested, human intervention is necessary for final touches and filtering out unusable bundles of response options.

Further evidence is needed to validate the utility of our method. Anticipated future work includes qualitative evaluation of the response clusters from education professionals, as well as psychometric evaluation to ensure that parallel items are of similar empirical difficulty. However, we believe the proposed method for empirically generating of MC response options illustrates how technology can be a practical aid to educators.

**Figure 1**

*2D visualization of hierarchical clusters of student responses. Example responses (some truncated) for each cluster are included as well as the expert rating levels from 0 to 3.*




## References

Burton, S. J., Sudweeks, R. R., Merrill, P. F., & Wood, B. (1990). *How to prepare better multiple-choice test items: Guidelines for university faculty* (Doctoral dissertation, Brigham Young University. Department of Instructional Science).

Gierl, M. J., Lai, H., & Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items. *Medical education*, *46*(8), 757-765.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, *15*(3), 309-333.

Nathan, M. J., & Petrosino, A. (2003). Expert blind spot among preservice teachers. *American educational research journal*, *40*(4), 905-928.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Zimmaro, D. M. (2004). Writing good multiple-choice exams. *Measurement and Evaluation Center. University of Texas at Austin. Web site*, *16*.

 © ISLS