Research Methods in Applied Linguistics 2 (2023) 100047

Contents lists available at ScienceDirect

Research Methods in Applied Linguistics

journal homepage: www.elsevier.com/locate/rmal

Investigating test content structure using multidimensional scaling

Abdolvahab Khademi

University of Massachusetts Amherst, United States

ARTICLE INFO

Keywords:
Multidimensional scaling
Dimension reduction
Construct validation
Writing assessment

ABSTRACT

Although multidimensional scaling (MDS) has extensively been used in social and physical sciences to visualize and explore the latent dimensionality of observed data, few studies in language assessment have used MDS for the purpose of test construction, including construct and content validation. In this study, we use MDS to investigate and compare the similarity of writing prompts in the IELTS and TOEFL iBT tests. Random prompts from both tests were presented as item pairs and raters were asked to rate the similarity of the pairs in terms of content and cognitive complexity. The results showed the writing prompts in the TOEFL iBT test represented two major dimensions while those in the IELTS test demonstrated three domains.

The need for measuring the construct validity of a test or instrument has resulted in a plethora of mathematical and statistical methods collectively known as dimension reduction techniques the purpose of which is to reduce the high dimensional data (e.g., data derived from many items, variates, or features) to a fewer number of theoretically and mathematically sound and representative underlying constituents or relationships that form the core(s) of the construct on which high-dimensional observations regress (the latent variables). Common multivariate statistical and mathematical methods used in measurement, science, and machine learning research include factor analysis, structural equation modeling, principal components analysis, singular value decomposition, linear discriminant analysis, t-distributed stochastic neighbor embedding (t-SNE), uniform manifold approximation and projection (UMAP), and multidimensional scaling.

Multidimensional scaling (MDS) is a general and intuitively appealing dimension reduction method that has found wide applications in fields as diverse as cognitive science, social sciences, developmental psychology, clinical psychology, psychophysics, neuroscience, marketing, political science, sociology, ecology, and linguistics, among others. Various studies in such fields that have used MDS include Russel and Bullock (1985), Pedelty, Levine, and Shevall (1985), Hornberger, Bell, Graham, and Rogers (2009), Hollins, Faldowski, Rao, and Young (1993), Youngentob, Johnson, Leon, Sheehe, and Kent (2006), Carroll and Green (1997), Amato (1990), Tu, Yan, Li, and Watts (2019), Lacher and O'Donnell (1988), Harnsberger (2001), Marozeau and McKay (2016), Machado, Duarte, and Duarte (2011), Machado and Eugénia Mata (2013), Kenkel and Orlóci (1986), and Meyer and Reynolds (2022). However, as far as the literature review in language assessment and applied linguistics research shows, multidimensional scaling has not been used as a dimension reduction procedure or exploratory structure analysis as commonly as other methods, such as factor analysis and structural equation modeling, in language assessment research and practice probably due to the fact that MDS as a dimension reduction method and validation procedure is not introduced to researchers in the language assessment field in the research and quantitative methods curriculum. This is despite the potential power and intuitive procedure of MDS, as shown in foregoing research studies in other fields but with similar purpose (dimension reduction). The present paper attempts to illustrate the use of MDS in language assessment and applied linguistics research through an empirical study. We begin by introducing an overview of the MDS method followed by its mathematical formulation, data considerations, and interpretation of results. Next, a review of literature follows in which those studies that have used MDS in language testing, education, or psychometrics will be presented. Finally, we will demonstrate a novel application of MDS for content validation in writing assessment.

E-mail address: vahab.khademi@gmail.com

https://doi.org/10.1016/j.rmal.2023.100047
Received 6 September 2022; Received in revised form 25 February 2023; Accepted 25 February 2023

An overview of multidimensional scaling

Multidimensional scaling (MDS) is a mathematical data reduction method that maps the distance (proximity) between observed data obtained from high dimensional space into distances in lower dimensional space (reducing the dimensions). The distance in the final MDS solution approximates the interrelation of the observed data items (objects). In other words, MDS shows the interrelationship of data (measured directly or indirectly) in a geometric space where the distance between points corresponds to the degree of relationship between observed data. In MDS mapping, small values of relationships map onto large distances between the points in the geometrical space and larger values of relationships are mapped onto small distances between the points in the geometrical space (Borg, Groenen & Mair, 2018). Interpreted the other way round, the closer the two points are in the MDS solution, the higher the correlation they represent of the variables (Borg et al, 2018).

MDS can be explained intuitively through this simple and concrete example from Izenman (2013). Imagine we have airline distance measurements among these 18 cities throughout the globe: Beijing, Cape Town, Hong Kong, Honolulu, London, Melbourne, Mexico City, Montreal, Moscow, New Delhi, New York, Paris, Rio de Janeiro, Rome, San Francisco, Singapore, Stockholm, and Tokyo. If a person has travelled to all these cities, we can ask them to tell which places they have visited, and they will name 18 cities (high dimension data). Is there another but more succinct (low dimensional) way for the traveler to tell us where they have travelled to? If we arrange the distances between those 18 cities, we will get a symmetrical 18×18 matrix, with half of the data cell values duplicates (because the distance between city A and B is the same as the distance between city B and city A) and the diagonal all zeros (because the distance between city A and city A is zero). Next, we use an algorithm from the MDS family to dig deeper into the relationship (distances) between those cities. The MDS results suggest that some cities are closer to each other and can form a meaningful cluster or dimension. For example, these cities lie close to each other and can make a cluster: {Rome, Paris, London, Moscow, Stockholm}, {Montreal, New York, Mexico City, San Francisco}, {New Delhi, Singapore, Hong Kong, Beijing, Melbourne, Tokyo} and {Rio de Janeiro}. We notice that we have discovered three dimensions and therefore can plot the cities (data points or objects) in a 3D plot. If we rotate and flip the plot (similar to the concept of rotation in factor analysis), we can map the cities to three continents. So, the traveler has travelled to three continents (three dimensions extracted from 18 high dimensional data points). This is a general description of the algorithms in MDS family of data reduction methods. As this example illustrates, MDS analyzes the distance between data points (objects) and discovers the original optimal latent structure that created those distances between the data points. As another example, suppose a researcher in applied linguistics is interested to evaluate the cultural sensitivity of a language proficiency test items among test taker populations, native culture speakers, and test developers (i.e., whether cultural sensitivity differs across groups). Because cultural sensitivity of items may be perceived differently by these groups, the researchers can conduct an MDS study and ask the different groups to rate the cultural sensitivity of items. If the solution results in a multidimensional configuration, the researcher may conclude that the items are not perceived homogenously in terms of cultural sensitivity, and they need to be reexamined to ameliorate possible issues. As another example application of MDS in applied linguistics, a researcher could wonder if perception of speaking topics differed across linguistic or cultural backgrounds. For instance, one would hypothesize that Germanic language native speakers would perform better on a speaking test than Asian speaking test takers. Through an MDS analysis, differential group functioning would be present if such hypothesis is established (see Chalhoub-Deville, 2016, for an example of MDS study on the perception of L2 oral proficiency construct compared across different backgrounds of native speakers). As another application of MDS in applied linguistics (language testing), suppose a researcher is interested in knowing the distribution of the difficulty of items in a reading comprehension test and if the actual difficulty data match the difficulty perceived by item developers. An MDS analysis solution will produce a likely multidimensional solution, in which each dimension will correspond to a difficulty level. The researcher can then examine the items that belong to each dimension and then investigate if they match the item difficulty as delineated by the item developers in the test blueprint.

The types of data used in an MDS analysis can generally be categorized as direct and indirect data (e.g., similarity ratings as direct data and correlations and co-occurrence data are indirect types of data). Direct data are collected by primary measurement of the relationship between or among the objects/items. Such data could emanate from methods of data collection such as sorting, ranking, and dissimilarity comparisons. In the direct approach to MDS data collection, participants directly judge the similarity of two items on a rating scale, such as the Likert scale. In the indirect approach to MDS data collection, proxy measures are used to infer similarity or distance, such as the time a participant spends in reacting to two stimuli or confusion data (Hout, Godwin, Fitzsimmons, Robbins, Menneer, & Goldinger, 2016).

With respect to the measurement scale, there are two families of MDS methods: metric and non-metric MDS. Metric MDS assumes that the data are quantitative and placed on a ratio scale and that there is a functional monotonic mapping relationship between the distance between the data points and the dissimilarities. Non-metric MDS assumes that the data is qualitative and has ordinal importance in which the purpose is to retain the ranks of the dissimilarities. In sum, metric MDS algorithms use ratio scale data, while nonmetric MDS algorithms use qualitative ranking information, that could be ordinal, interval, or ratio scaled (Tsogo et al., 2000).

Because MDS is a data reduction technique, one may compare it to other prevalent data reduction methods in mathematics and statistics, such as principal component analysis (PCA), factor analysis (FA) and isomap. Compared to PCA, FA, and isomap, MDS is more intuitive and has gained more application areas. In certain studies, MDS is preferred over FA because MDS does not require linearity and normality assumptions (MDS is primarily a mathematical model; however, probabilistic extensions have also been developed, as in Zinnes and MacKey, (1983), Oh and Raftery, (2001), and Sato-Ilic, (2020)). The only assumption of MDS is that the number of dimensions should be one less than the number of data items, which implies at least three variables and two dimensions (Saeed, Nam, Imtiaz Ul Haq, & Bhatti, 2018). Compared to FA, MDS has been mostly used in exploratory inquiries, hence fewer technical assumptions are needed (Tucker-Drob & Salthouse, 2009). The main difference between PCA and MDS is that PCA uses

covariance data while MDS uses direct or indirect similarity data. However, if MDS is treated as classical scaling, it becomes an eigendecomposition problem and is the same as PCA (Izenman, 2008). A main difference between FA and MDS is that in MDS a general factor has no effects, which occurs when items are highly correlated (Ding, 2018).

There are several computer programs that can perform MDS analysis, the most common of which are PROXSCAL, SYSTAT, R, the SMACOF package in R (De Leeuw & Mair, 2009), SAS, Matlab, and IBM SPSS. This study uses IBM SPSS to perform MDS analysis on the collected data.

Mathematics of multidimensional scaling

Multidimensional scaling is a mathematical data reduction and data visualization technique that makes use of the geometrical distance between data points which then creates a reduced map that can represent the original space. In doing so, MDS finds and converts dissimilarity data to geometrical data between pairs of data objects. MDS computes the distances among the data points through a distance function d_{ij} which is used to compute the distance between items i and j ($i \neq j$). There are different mathematical functions to measure the distance between values and vectors in a metric space. One most common metric space in MDS applications is the Euclidean space, where the Cartesian axes represent the dimensions or principal axes to visualize the emergent latent structure in the data. The Euclidean distance d_{ij} between objects i and j (i ≠ j) in the m-dimensional Euclidean space $\mathbf{X}$ is calculated as:

$$d_{ij}(X) = \sqrt{\left(x_{i1} - x_{j1}\right)^2 + \ldots + \left(x_{im} - x_{jm}\right)^2} \tag{1}$$

$$= \left(\sum_{a=1}^{m} \left(x_{ia} - x_{ja}\right)^2\right)^{1/2} \tag{2}$$

In Equation (1) and Equation (2), x_{ia} and x_{ja} represent the values of points i and j on coordinate axis a, respectively. After the distances are measured through a distance function (e.g., the Euclidean function or another variant of the Minkowski distance), the MDS model is fit through minimizing a loss function (this is similar to the least-squares optimization method). The MDS loss function measures the difference between the observed data and the predicted values ($\hat{d}$). The sum of these model fit errors is called the raw Stress (σ^2_{raw}), calculated through Equation (3):

$$\sigma^2_{raw} = \sum_{i<j} e^2_{ij} = \sum_{i<j} \left(f\left(\delta_{ij}\right) - d_{ij}(\mathrm{X})\right)^2 \tag{3}$$

In Equation (3), δ_{ij} are the dissimilarity values (because in MDS, the preference is to input dissimilarity data and often the observed similarity data are rescaled as dissimilarity data), $f(\delta_{ij})$ is the regression function that produces disparities (which is similar to predicted values in ordinary least square regression), $\hat{d}_{ij}$ are the predicted values, and $d_{ij}(\mathrm{X})$ represent the Euclidean distances calculated using Equation (1) above. The raw Stress loss function is usually normalized by dividing over the sum of the squared distances, giving the Stress-1 loss index (Borg et al, 2018) in Equation (4):

$$Stress\ 1 = \sqrt{\sum_{i<j} \left(\hat{d}_{ij} - d_{ij}(X)\right)^2 / \sum_{i<j} d^2_{ij}(X)} \tag{4}$$

When the raw Stress value is normalized, the goodness of fit can be compared across different MDS solutions (Borg et al., 2018), for instance, when comparing models with different number of dimensions. As the number of dimensions increases in the MDS solution, the Stress value falls. Overall, the value of the Stress changes as a function of different factors, most important of which are the number of points (n), the dimensionality of the MDS solution (m), the error component of the data (i.e., model fit), the family and type of the MDS model applied, the existence of outliers, the number of ties in the points (in the nonmetric MDS), and the proportion of missing data (Borg et al., 2018).

Data considerations in MDS

The types of data used in an MDS analysis can be broadly categorized as direct and indirect data. Direct data are collected by primary measurement of the relationship between or among the objects/items. Such data could come from methods of data collection, such as sorting, ranking, and item comparisons. The most common type of comparison data is pair-comparison, in which the participant is asked to rate the similarity/dissimilarity of items i and j on a Likert scale (e.g., 1 = “Not similar at all”, 8 = “Highly similar”, on an eight-point Likert scale). In pair-comparisons, the number of pairs may increase fast as more items are added (number of pairs = $n(n-1)/2$). In such cases, some pairs can be removed and regarded as missing data in one matrix, but as valid data in another matrix (Jaworska et al., 2009) without invalidating the results. In fact, simulation studies show that up to 30% missing data will not affect the results of an MDS analysis (Rosett & Klein, 1995). In pair comparisons, when the number of items becomes large, rating may become tedious, inaccurate, and inefficient, warranting for alternative methods, including grouping, ranking, and their subclasses. Sorting, grouping, and incomplete data analysis are alternatives to pair comparison which are briefly described below.

Sorting data. In sorting or grouping methods of data collection, m participants are asked to produce k groups into which similar items are placed. Once the sorting task is completed, there will be an $n \times n$ incidence matrix per participant the entries of which are either 1 (if the row and column items belong to the same group) or 0 (if not). Finally, these matrices are summed to produce

a similarity matrix **F** the entries of which are the frequency of occurrence of each item in the same group (Tsogo et al., 2000). In the hierarchical sorting methods, the participant judges the items to be similar and also the groups to be similar. Similar groups are merged together subsequently to form a single group that contains similar items. The dissimilarity of two objects is defined as the number of disjoint groups.

Ranking data. In the ranking method of creating input data for MDS analysis, an item is chosen as a reference and the rest of items are ranked from most to least similar to the reference item. The reference item is iterated across all the items.

Incomplete data. When the number of dissimilarities becomes prohibitively large, using incomplete similarity tasks that produce incomplete matrices is one practical and efficient solution. Incompleteness can be introduced into the data collection and the input data in random or cyclic graph models. In the random graph model, items are considered as the vertices and their dissimilarities as edges. The edges are randomly (and based on an appropriate proportion value) dropped. In the cyclic deletion graph model, edges are deleted in an iterative manner such that in each iteration, the graph is cyclic, regular, and connected (Tsogo et al., 2000). Spence and Domoney (1974) in a Monte Carlo simulation study showed that a cyclic or random incomplete dissimilarity matrix with one-third of values deleted satisfactorily recovered the geometrical structure of item relationships.

Interpretation of MDS results

Depending on the purpose of the MDS analysis, the results may need subjective interpretation as for the number and, in particular, the meaning of the dimensions or the clusters of data derived, similar to the interpretation of the derived factors in exploratory factor analysis or other dimension reduction procedures. This step of MDS requires domain knowledge.

Davison and Sireci (2000) suggest three criteria in determining and interpreting the dimensionality structure of MDS analysis results: model fit, interpretability of the solution, and reproducibility (replication). The latter two go hand-in-hand. The model fit is evaluated using the STRESS (PROXSCAL) or S-STRESS (ALSCAL) values (the lower and closer to zero, the better fit), which indicate the difference between the scaled distances and transformed proximities (Davison & Sireci, 2000). R-squared (RSQ) is another index which shows the proportion of variance in the distances accounted for by the transformed data. For the sake of parsimony, the researcher needs to determine the lowest dimensionality which fully represents all aspects of the data structure. In addition to the visual inspection of the dimensions, Kruskal and Wish (1978) also recommend interpreting the neighborhoods formed along the dimensions, especially when MDS gives a high-dimensional solution.

The results of an MDS study can be combined with other computational methods to investigate similarity research studies. For instance, Hout, Godwin, Fitzsimmons, Robbins, Menneer, and Goldinger, (2016) applied MDS to visual search data and then combined the results with behavioral measurements and eye-tracking data from participants.

A survey of previous studies in testing and assessment using MDS

MDS has extensively been used in data visualization and structural hypotheses evaluation in many scientific and social science studies, particularly in psychology. In structural hypothesis evaluation, the assumption is that points in psychological space are spanned by the items' subjective attributes, and by running similarity judgments the distance between the points in the psychological space can be measured (Borg et al., 2018). Because the purpose of the present paper is to investigate structural hypothesis in language testing, first a survey of studies in assessment that have used MDS as an analysis tool will be presented to provide a glimpse of MDS applications, types of research questions that can be addressed by MDS in language testing and educational measurement, and how the results from an MDS analysis are interpreted.

A number of studies and use cases have applied MDS in test construction and validation, including item analysis with respect to the number of dimensions the items load on (Jaworska & Chupetlovska-Anastasova, 2009). A few studies using MDS in applied linguistics, education, and psychological assessment are presented and summarized below.

Oltman, Stricker, and Barrows (1990) used weighted multidimensional scaling to determine the latent structure of the TEOFL test and if the test takers' language background (seven languages) and proficiency level (three levels; 21 groups in total) affected the dimensionality of the factor structure of the test. The authors converted actual responses (from 53,169 test responses) to 146 items into tau association (an example of indirect type of data used in MDS) and produced a 146×146 matrix of tau correlations between the items. In addition, the analysis produced a 21×21 weight matrix for the 21 groups of test takers to assess if groups differed in perception of the TOEFL factor structure using SINDSCAL procedure (Pruzansky, 1975). As for the structure of the TOEFL test, the authors selected a four-dimensional solution (corresponding to the three sections of the TEOFL test, and one additional dimension accounting for difficult items in the reading passages). The results for the group variations in test structure showed that perception of the factor structure was homogenous across language groups but differed across proficiency groups (low-scoring examinees perceived more distinct dimensions in the SINDSCAL solution).

In another study, Sireci and Geisinger (1995) applied MDS to explore the content validity of two tests to support content validation of the instruments in comparison with the hypothesis posed in the test blueprints (similar to confirmatory factor analysis). In their study, Sireci et al. (1995) used two groups of 15 subject matter experts (SME) with at least three years of domain experience to rate the similarity of pairs of items from the test. The researchers in that study performed multidimensional scaling analysis using the INDSCAL algorithm (Carroll & Chang, 1970). The INDSCAL algorithm is used when the raters' contributions (i.e., weights) are also taken into the account or investigated. The results of their study showed that MDS analysis was able to reveal the relationship between clusters of items that well represented the content structure hypothesized in the original test blueprint.

In a similar study on content structure, Sireci and Geisinger (1992) used MDS analysis to assess the mapping of the content domains specified in a 30-item test blueprint onto the structure manifested in similarity ratings of the items. In that study, the researchers utilized three experienced content domain raters to rate the similarity of items in a study skills test to explore the perceived content domain structure by the raters. The raters were unaware of the number of content domains in the test blueprint and were not directed to the aspect of rating. The results showed a close match between the MDS solution and the domains hypothesized in the test blueprint.

Another empirical study that used MDS to test structural hypothesis was that of Borg, Bardi, and Schwartz (2017), where the researchers administered a 40-item questionnaire that measured the perceived importance of ten basic personal values to 151 adults who rated on a six-point scale the extent to which the portrayal of person in each item resembled themselves. The purpose of the study was to evaluate the structural hypothesis that groupings of items represent the ten personal values purported in the questionnaire. The results showed that the emergent circular pattern matched the structural hypothesis and also previous replications by other researchers.

Studying teachers' perception of English language learners in the United States, Alvarez (2019) collected 400 teacher statements about English learners and performed an MDS on these data to explore what teachers perceived on English learners. The 400 statements were sorted into 28 categories that generated 92 data items that were to be sorted by 40 teachers (20 pre-service and 20 in-service) in a historically white dominant institution. The findings of their study showed that teachers' perceptions were influenced by their teaching status (pre-service versus in-service).

In another application of MDS, O'Donnell and Sireci (2021) applied multidimensional scaling to assess the perception of achievement level labels by teachers and parents in statewide testing programs. The authors administered sorting data collection method to rate similarity of labels. The authors found that teachers and parents differed in their perception of labels that denoted the same level of performance achievement.

Present study: comparing the content structure of IELTS and TOEFL writing prompts

MDS is a mathematical tool that has been used in the process of test construction and validation, similar to factor analysis models. In this section of the paper, we present a research study in which MDS was used to answer the research question: whether or not the contents in the writing prompts of IELTS and TOEFL iBT are similar in their structure and complexity and to what extent they measure writing construct in an academic setting.

In this application of MDS in language assessment, we posed the research question about the similarity of the content and structural dimensions of the writing prompts of the Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS), developed and administered by ETS and Cambridge ESOL, respectively. International students aspiring to study in one of the major English-speaking countries, such as in North America, Europe, and Australia, are required to demonstrate adequate academic English proficiency by participating in one of these two proficiency examinations and achieve an institution-specified minimum score.

The writing skill component of these test batteries provides a measure of readiness for academic writing using prompts that simulate settings in which the prospective student will be required to write academic essays. In the present study, we were interested in investigating the number of dimensions or domains that the set of prompts from each examination exposed test takers to and what those domains represented in terms of task-based language testing. The assumption was that both writing components of these two examinations should measure content related to academic settings or related to the academic writing proficiency construct. Investigating the content structure of writing prompts is a test validation process because a test should assess second language abilities independent of general cognitive abilities (Chapelle, Enright & Jamieson, 2010) and other construct-irrelevant factors, such as task complexity, task specificity, task familiarity or cultural bias. Validity is defined as, "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Therefore, in assessing the writing proficiency of students, selection of prompts with no or minimum construct-irrelevant factors plays an important role in the interpretation and use of the test scores obtained.

The data in the present study comprised pair ratings of writing prompts from official TOEFL (Educational Testing Services, 2009) and IELTS test papers released by ETS and Cambridge ESOL. For this study, 40 prompts (20 prompts from the TOEFL iBT writing test and 20 prompts from IELTS Academic Writing Task II examination) were randomly chosen from the officially released pool of topics by ETS (in the Official Guide to TOEFL Test, 2009) and Cambridge University Press past examinations book series and paired with each other to form 780 pairs. Each pair was rated on a 10-point Likert scale in terms of cognitive complexity similarity of the prompts. Before starting the similarity rating process, raters were introduced to the concept of task complexity with an example and instruction for rating. Lohman and Lakin (2011) define task complexity as the number of cognitive processes in a task, the importance of the cognitive processes, demands exerted on attention and memory, and adaptation and executive functions required. Each item pair required the raters to rate how similar the pair of prompts were in terms of cognitive complexity for the students. In the introduction section of the questionnaire, task complexity was defined as the amount of thinking, background knowledge, and experience that may be needed by a student to write about a topic in a testing setting. Raters were instructed about task complexity and how to rate the paired prompts at the beginning of the survey, which can be summarized as, "Students may be challenged to develop a writing prompt by the cognitive difficulty of the prompt, including aspects such as abstract thinking, topical knowledge, rather than the linguistic resources they need to write about those prompts. In this survey, please rate how similar two prompts are to each other in terms of cognitive demand they exert on the test taker to develop." (See Appendix A for the full text of the rating instruction used in the survey.)

Table 1
Model Fit Indices for the Initial Run of the PROXSCAL Algorithm for Six Dimensions for the TOEFL Writing Prompts.

Fit Index	Dimensions				
	2	3	4	5	6
S-Stress	0.24	0.17	0.13	0.10	0.09
RSQ	0.68	0.71	0.74	0.78	0.81

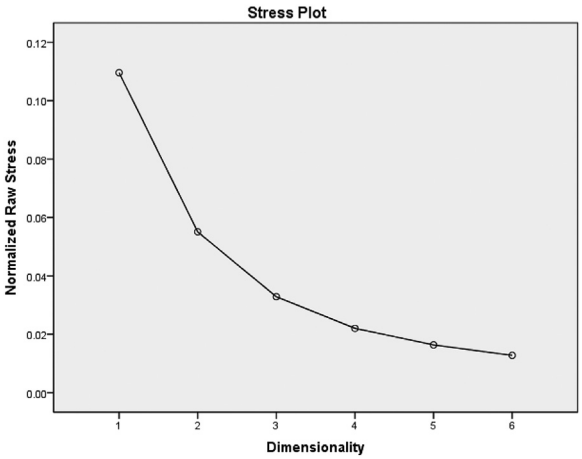

Fig. 1. Scree Plot from PROXSCAL MDS Analysis of the TEOFL Prompts Data.

Rating was conducted through online surveys completed by three raters. The raters were non-native speakers who had formal education in language teaching and applied linguistics at the master's and doctoral levels with a minimum of 10 years teaching different language skills. The raters were teachers of IELTS and TOEFL at the time of the present study and resided in North America. Two of the raters were female and one was male. The raters read both prompts in pairs and decided if they were very different or very similar in terms of the cognitive complexity of the tasks (according to the instruction). Because the number of item pairs was large in this study, the rating process proceeded in multiple sessions to address rater fatigue and time constraints.

The collected data were entered in IBM SPSS v. 19 in the form of three 40 × 40 symmetrical lower triangle matrices (this paper, though, reports on the results from three 20 × 20 matrices separated from the original parent matrix for IELTS and TOEFL iBT data). Because there are three similarity matrices, the MDS method used is called replicated MDS, or RMDS. Both PROXSCAL and ALSCAL algorithms were used to scale the data. The level of measurement chosen was ordinal with ties broken. In addition to ALSCAL procedure, PROXSCAL was also used to aid the interpretation of the obtained dimensions. In PROXSCAl, the initial configuration was changed from simplex to Torgerson, as recommended by Borg et al. (2018), which helps avoid suboptimal local minima.

Table 1 shows the results of the ALSCAL analysis on the ordinal data of the TOEFL writing topics. The common indicators for goodness-of-fit are the loss function S-STRESS and the R squared (RSQ) values. The lower the S-STRESS is, the better the solution. For RSQ, we're seeking higher values.

Similar to factor analysis, MDS analysis will also output a scree plot to visually help determination of the number of dimensions in the dataset. Figure 1 below shows the results of the analysis using the PROXSCAL algorithm in the form of a scree plot.

As the stress plot suggests, two to three dimensions are likely in the structure of TOEFL writing prompts.

As both the ALSCAL and PROXSCAL procedures showed, a two-dimensional interpretation is possible considering the SSTRESS value and the RSQ and the interpretability of the solution. Figure 2 shows the two-dimensional structure for the TOEFL writing prompts. The axes are rotated for better interpretation of the dimensions.

If we rotate the axis counterclockwise about 23 degrees, we can see that two coordinates of items are formed: T03, T04, T15, T19, T13; T10, T12, T16, T07, T08 and T14 on one dimension, and T02, T06, T11, T05; T09, T01, T18 and T20, on the other dimension. Now we can study these prompts and see what characteristic(s) distinguishes them. What we see is that in the first group of prompts, the examinee is required to write about something personal and known through personal experience or an experience that has affected their "self". The second cluster of this dimension asks examinees to write about something distal from their personal experience; it is about society, environment or experiences to be gained in the future (e.g., running one's own business). We may designate this dimension Personal and Societal Development. The second dimension is trickier to interpret. Nevertheless, it can be seen that most prompts on this dimension pivot around evaluating the structures or components of a modern society in terms of their utility and justification, such as attending classes, the use of a zoo and spending money on space projects.

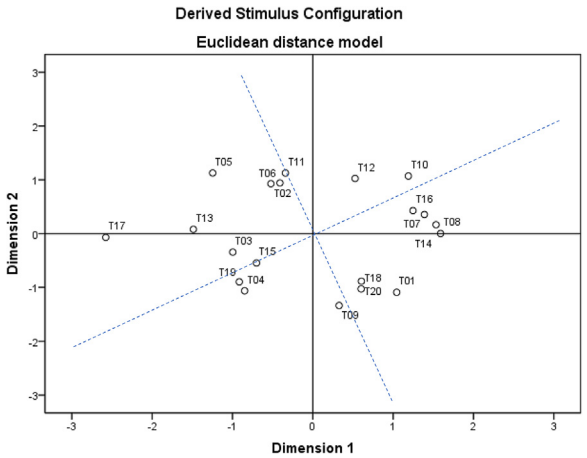

Fig. 2. Two-dimensional Rotated Structure for the TEOFL Prompts Data.

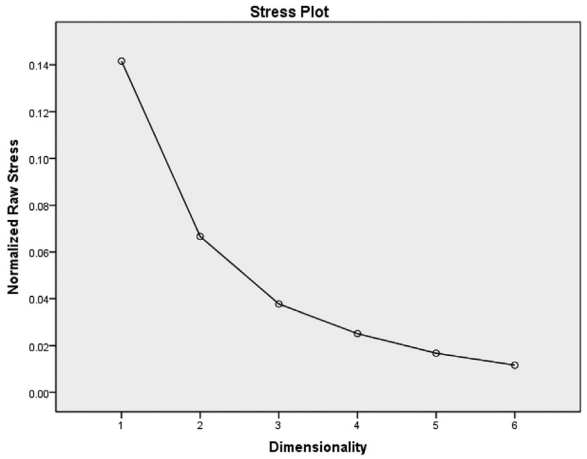

Fig. 3. Scree Plot from PROXSCAL MDS Analysis of the IELTS Prompts Data.

Table 2
Model Fit Indices for the Initial Run of the PROXSCAL Algorithm for Six Dimensions for the IELTS Writing Prompts.

Fit Index	Dimensions				
	2	3	4	5	6
S-Stress	0.29	0.21	0.16	0.13	0.11
RSQ	0.51	0.56	0.58	0.63	0.65

Table 2 shows the results of ALSCAL analysis on the ordinal data of the IELTS writing prompts. The common indicators for goodness-of-fit are the loss function S-STRESS and the R squared (RSQ) values. The lower the S-STRESS is, the better the solution. For RSQ, we're seeking higher values.

Figure 3 shows the results of the analysis using the PROXSCAL algorithm in the form of a scree plot.

The PROXSCAL scree plot shows a two- or a three-dimensional structure to the IELTS writing prompts.

As both ALSCAL and PROXSCAL procedures showed, a three-dimensional interpretation is possible considering the STRESS value and the RSQ and the interpretability of the solution. Figure 4 show the three-dimensional structure for the IELTS writing prompts.

The best way to identify the dimensions in 3D space is to rotate the graph in different angles and perspectives. In doing so, prompts 1, 2, 4, 6, 7, 14 and 17 fall on one dimension; prompts (2), (5), 9, 10, 11, 12, 13, (17), and 18 fall on the second dimension; and prompts 3, 5, 8, (14), 15, 16, 19, and 20 fall on the third dimension (prompts in parentheses are gravitated to more than one dimension). What we discern in the first set of prompts in the IELTS test is the dominant topics related to the individual and their interaction with another individual or the society at large, such as motherhood, community service and children growing into social

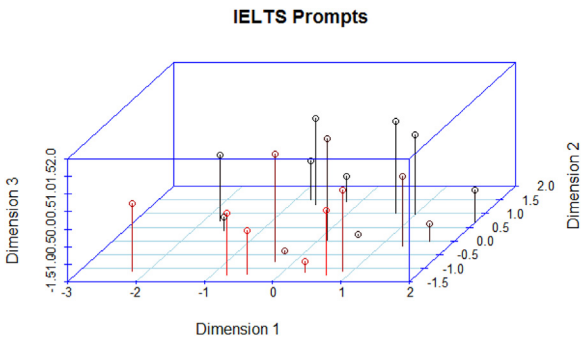

Fig. 4. Three-dimensional Rotated Structure for the IELTS Prompts Data.

adults. The second dimension mostly includes topics about analyzing the issues of the society, such as salary disparity, effect of technology on relationships, and health issues. The third dimension also threads social issues together, but regarding issues caused by social structure and not by social behaviors, such as censorship, capital punishment and international relationships. The second and third dimensions are similar to each other and probably constitute a single dimension conceptually.

Discussion

To what extent test items measure the construct or the content is a matter of test validation and score interpretation. An item must be clear of irrelevant constructs or factors for the score interpretation to be valid. This study differs from a typical test validation study in that the items (writing prompts) in this study never appear together as a test. Nevertheless, they do belong to the same testing purpose: assessing the writing proficiency of non-English speaking candidates. Thus, performing a validation study on the prompts to find different dimensions or categories of items is justified. Finding different dimensions or categories of prompts helps comparability of test scores and the appropriateness of the prompts for different examinee groups, in terms of education, age and cognitive attributes.

In both the TOEFL and IELTS writing prompts, one category of the items asks examinees to write about personal and social experiences, which are more immediate and concrete to them. However, in a second category, the examinees in both tests are required to think more abstractly as problem solvers or judges regarding the social issues and problems. The IELTS prompts add the complexity by projecting a third dimension of detached factors affecting the social structure, which is more abstract and requires higher cognitive ability to resolve. These issues may pose preferential selection of prompts for different examinees with different education, age, and cognitive attributes.

Implications of the study

The results of the present study reflect implications both on the substantive and methodological aspects.

As for the former aspect, the results of the present study bear implications for test developers, test takers, and test score users. For test developers, the results show that writing prompts need to be aligned with different attributes of test takers or more and exhaustive choices be given to the test takers in selecting the prompts. Because test takers with different demographic backgrounds may respond differently to the same prompt, providing a pool of prompts or designing prompts where demographic backgrounds of test takers are taken into consideration will affect the test fairness in a positive way. In addition, test takers are reminded that writing proficiency is not just defined by knowing about the language, but rather it involves different levels of thinking and prior knowledge and experiences. Though this could be a counterargument to the construct validity of the test, in reality background knowledge and topical familiarity play nonnegligible role in responding to writing prompts. As for the test users, the implications regard determining cut-off scores for different tests and score comparability across different writing tests. If test takers with different demographic backgrounds attain different scores from a writing test (despite their similar writing proficiency level), then any decision making with respect to the scores on such tests should consider such demographic variability.

As for the implication on the methodological level, the present study has used a mathematically elegant model to explore the content dimensionality of writing prompts. The prominent feature of the methodological design of the present study is the formation of prompts as a cohesive test or questionnaire. Writing prompts are usually investigated for their validity and psychometric properties individually. Our method demonstrates how disjoint items can be grouped together to measure their similarity in terms of content structure or cognitive complexity.

Limitations of the study

Despite the acceptable results from the present MDS study on the content structure of the IELTS and TEOFL iBT writing prompts, the research could be improved by addressing some limitations of the present study.

Although the MDS method does not require a large number of participants (due to its deterministic nature), having more than three raters could increase the precision of the results and its generalizability. In the present study, we used three raters due to logistic constraints. Although the results are still valid, a higher number of raters with a more diverse background in education, experience, culture, and first language (native vs. non-native raters) would enrich the study. Nonetheless, the present study with its limited number of raters can serve as a method demonstration for future studies.

We used 40 items for this study, producing 780 similarity items pairs to be rated by the participants. This number is definitely very large and may have affected the raters in terms of their time and fatigue, although the rating process was conducted in multiple sessions. However, decreasing the number of items or using incomplete MDS procedure (either random or cyclic graph models) or using sorting methods could enhance the logistic efficiency of the study.

These limitations could be addressed in future studies by the authors or other interested researchers and compare the results with the present study.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A

Rating Instruction:
Dear rater,

The present rating task is aimed at comparing topics used in different writing tests taken by English language learners or those who need to demonstrate their written communicative proficiency in English. As experienced writing teachers or writing assessors, you know that some topics are more or less difficult for your students to develop. The difficulty does not necessarily lie in the language resources they need to develop the topic. Rather, the difficulty or complexity may lie in the cognitive demand exerted by the topics, which may require different amounts and levels of thinking, background knowledge, and experience by a student to write about that topic. For instance, given students are equally proficient in English, writing about “different ways to meet new friends” might be easier for some students than writing about “different ways the society shapes friendships”. The latter seems to be more demanding in terms of analyzing the topic, developing a thesis, stating arguments and examples to support the thesis and expressing a personal view on the thesis statement. So, in this case we can see that these two topics are not that similar in terms of cognitive complexity. As experienced writing teachers or writing assessors, you can determine if two writing topics are very similar to or very different from each other in terms of cognitive complexity. In fact, similarity of two topics can be rated on a continuum from 1 (Not similar) to 10 (Very Similar), with degrees in between. For instance, for the topics just mentioned above, I give a score of 4 for their low similarity:

Different ways to meet new friends.

○	○	○	○	○	○	○	○	○	○
1	2	3	4	5	6	7	8	9	10

Different ways the society shapes friendships.

So below you can see pairs of writing topics. Please rate each pair on the basis of similarity of the topics in terms of cognitive complexity you perceive. Please mark the score by selecting a value on the scale.

Thank you for your kind cooperation.

References

Alvarez, C. (2019). *Teachers’ perceptions of English learners: a multidimensional scaling approach*. Illinois State University [Unpublished doctoral dissertation].

Amato, P. R. (1990). Dimensions of the family environment as perceived by children: a multidimensional scaling analysis. *Journal of Marriage and the Family, 52*(3), 613–620.

American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education. (2014) Standards for educational and psychological testing. AERA.

Borg, I., Bardi, A., & Schwartz, S. (2017). Does the value circle exist within persons or only across persons? *Journal of Personality, 85*(2), 151–162.

Borg, I., Groenen, P. J. F., & Mair, P. (2018). *Applied multidimensional scaling and unfolding.* Springer.

Carroll, J. D., & Chang, J. J. (1970). An analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika, 35*(3), 238–319.

Chalhoub-Deville, M. (2016). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16–33. 10.1177/026553229501200102.

Carroll, J. D., & Green, P. E. (1997). Psychometric methods in marketing research: Part II, multidimensional scaling. *Journal of Marketing Research, 34*(2), 193–204.

Chapelle, C. A, Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3–13.

Davison, M. L., & Sireci, S. G (2000). Multidimensional scaling. In H. E. A. Tinsley, & S. D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling.* Academic Press.

De Leeuw, J., & Mair, P. (2009). Multidimensional scaling using majorization: SMAACOF in R. *Journal of Statistical Software, 31*(3), 1–30. http://www.jstatsoft.org/v31/i03/.

Ding, C. S. (2018). *Fundamentals of applied multidimensional scaling.* Springer.

Educational Testing Service. (2009). *The official guide to the TOEFL test.* McGraw-Hill.

Harnsberger, J. D. (2001). The perception of Malayalam nasal consonants by Marathi, Punjabi, Tamil, Oriya, Bengali, and American English listeners: A multidimensional scaling analysis. *Journal of Phonetics, 29*(3). 10.1006/jpho.2001.0140.
Hollins, M., Faldowski, R., Rao, S., & Young, F. (1993). Perceptual dimensions of tactile surface features: A multidimensional scaling analysis. *Perception & Psychophysics, 54*, 697–705.
Hornberger, M., Bell, B., Graham, K. S., & Rogers, T. T. (2009). Are judgements of semantic relatedness systematically impaired in Alzheimer's disease? *Neuropsychologia, 47*(14), 3084–3094.
Hout, M. C., Godwin, H. J, Fitzsimmons, G., Robbins, A., Menneer, T., & Goldinger, S. D (2016). Using multidimensional scaling to quantify similarity in visual search and beyond. *Attention, Perception, & Psychophysics, 78*, 3–20. 10.3758/s13414-015-1010-6.
Izenman, A. J. (2013). *Modern Multivariate Statistical Techniques: regression, classification, and manifold learning*. Springer.
Jaworska, N., & Chupetlovska-Anastasova, A. (2009). A review of multidimensional scaling (MDS) and its utility in various psychological domains. *Tutorials in Quantitative Methods for Psychology, 5*(1), 1–10.
Kenkel, N. C., & &Orlóci, L (1986). Applying metric and nonmetric multidimensional scaling to ecological studies: Some new results. *Ecology, 67*(4), 919–928.
Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Sage.
Lacher, D. A., & O'Donnell, E. D (1988). Comparison of multidimensional scaling and principal component analysis of interspecific variation in bacteria. *Annals of Clinical and Laboratory Science, 18*(6), 455–462.
Lohman, D., & Lakin, J. (2011). Intelligence and reasoning. In R. Stenberg, & S. Kaufman (Eds.), The Cambridge Handbook of Intelligence (pp. 419–441).
Machado, J. A. T., Duarte, G. M., & Duarte, F. B. (2011). Analysis of financial data series using fractional Fourier transform and multidimensional scaling. *Nonlinear Dynamics, 65*(3), 235–245.
Machado, J. A. T., & Eugénia Mata, M (2013). Multidimensional scaling analysis of the dynamics of a country economy. *The Scientific World Journal, 2013*. 10.1155/2013/594587.
Marozeau, J., & McKay, C. M. (2016). Perceptual spaces induced by cochlear implant all-polar stimulation mode. *Trends in Hearing, 20*. 10.1177/2331216516659251.
Meyer, E. M., & Reynolds, M. R. (2022). Multidimensional scaling of cognitive ability and academic achievement scores. *Journal of Intelligence, 10*(4), 117. 10.3390/jintelligence10040117.
O'Donnell, F., & Sireci, S. G. (2021). Language matters: teacher and parent perceptions of achievement labels from educational tests. *Educational Assessment, 27*(1). 10.1080/10627197.2021.2016388.
Oh, M. S., & Raftery, A. E. (2001). Bayesian multidimensional scaling and choice of dimension. *Journal of the American Statistical Association, 96*(455), 1031–1044. 10.1198/016214501753208690.
Oltman, K. P., Stricker, L., & Barrows, T. S. (1990). Analyzing test structure by multidimensional scaling. *Journal of Applied Psychology, 75*(1), 21–27.
Pedelty, L., Cohen, L. S., & Shevall, S. K. (1985). Developmental changes in face processing: results from multidimensional scaling. *Journal of Experimental Child Psychology, 39*(3), 421–436.
Pruzansky, S. (1975). *How to use SINDSCAL-A computer program for individual differences in multidimensional scaling*. Bell Telephone Laboratories Unpublished report.
Rosett, T. R., & Klein, B. P. (1995). Efficiency of a cyclic design and a multidimensional scaling sensory analysis technique in the study of salt taste. *Journal of Sensory Studies, 10*(1), 25–44.
Russel, J. A., & Bullock, M. (1985). Multidimensional scaling of emotional facial expressions: similarity from pre-schoolers to adults. *Journal of Personality and Social Psychology, 48*(5), 1290–1298.
Saeed, N., Nam, H, Imtiaz Ul Haq, M., & Bhatti, D. (2018). A survey on multidimensional scaling. *ACM Computing Surveys, 51*(3). 10.1145/3178155.
Sato-Ilic, M. (2020). Probabilistic metric based multidimensional scaling. *Procedia Computer Science, 168*, 65–72. 10.1016/j.procs.2020.02.258.
Sireci, G. S., & Geisinger, K. F. (1995). Using subject-matter experts to assess content representation: an MDS analysis. *Applied Psychological Measurement, 19*(3), 241–255.
Spence, I., & Domoney, D. (1974). Single subject incomplete designs for nonmetric multidimensional scaling. *Psychometrika, 39*(4), 469–490.
Tsogo, L., Masson, M. H., & Bardot, A. (2000). Multidimensional scaling methods for many-object sets: A review. *Multivariate Behavioral Research, 35*(3), 307–319.
Tu, W., Yan, M., Li, Q., & Watts, J. (2019). Attitudes toward disabilities among students in college settings: A multidimensional scaling analysis with biplot. *The Australian Journal of Rehabilitation Counselling, 25*(2), 79–95. 10.1017/jrc.2019.10.
Tucker-Drob, E. M., & Salthouse, T. A. (2009). Confirmatory factor analysis and multidimensional scaling for construct validation of cognitive abilities. *International Journal of Behavioral Development, 33*(3), 277–285. 10.1177/0165025409104489.
Youngentob, S. L., Johnson, B. A., Leon, M., Sheehe, P. R., & Kent, P. F. (2006). Predicting odorant quality perceptions from multidimensional scaling of olfactory bulb glomerular activity patterns. *Behavioral Neuroscience, 120*(6), 1337–1345.
Zinnes, J. L., & MacKay, D. B. (1983). Probabilistic multidimensional scaling: complete and incomplete data. *Psychometrika, 48*, 27–48. 10.1007/BF02314675.