Language Testing in Asia

RESEARCH Open Access

C-Test construct validity: Evidence from nonparametric item response theory

Roya Shoahosseini[1], Purya Baghaei[2*], Hossein Khodabakhshzadeh[1] and Hamid Ashraf[1]

*Correspondence:
puryabaghaei@gmail.com

[1] Department of English, Torbat-e Heydarieh Branch, Islamic Azad University, Torbat-E Heydarieh, Iran
[2] Department of English, Mashhad Branch , Islamic Azad University, Mashhad, Iran

Abstract

C-Test is a gap-filling test designed to measure first and second language proficiency. Over the past four decades, researchers have shown the fit of C-Test data to parametric item response theory (IRT) models, but no study so far has shown the fit of C-Tests to nonparametric IRT models. The purpose of this study is to contribute to the ongoing C-Test validation project by providing evidence of fit to the Mokken scale analysis as a widely used nonparametric IRT model. A six-passage C-Test battery was analyzed using the monotone homogeneity model and the double monotonicity model of Mokken. Unidimensionality was evaluated using the automatic item selection procedure. Findings showed that the C-Test passages form a strong unidimensional scale, fit well to the monotone homogeneity model, and, after deleting one item, have a sufficient fit to the double monotonicity model. The findings also indicated that the items form a hierarchy, and persons can be located on an ordinal scale using their C-Test sum scores. Implications of the study for C-Test validity and application are discussed.

Keywords: C-Test, Nonparametric IRT models, Mokken scaling, Monotone homogeneity model, Double monotonicity model

Introduction

C-Test is a gap-filling test designed to measure language proficiency in both first and second language (Klein-Braley, 1985). A C-Test is composed of four to eight short independent texts in which the second half of every second words is deleted. Deletions start from the second sentence. The first and the last sentences in each passage remain intact to provide some context for text processing. Examinees have to fill in the missing letters. Raatz and Klein-Braley (1982) proposed the C-Test as an improvement over the classical cloze test. They argued that cloze test suffers from some problems including change in the test's psychometric qualities with the change of onset of deletions and deletion rates. Several studies showed that C-Tests with different points of onset of deletions and different rates of deletions have different difficulty levels and have different correlation coefficients with external criteria (Alderson, 1983; Raatz & Klein-Braley, 1982). By fixing the rate and the point of onset of deletions, Klein-Braley aimed at a more stable construct for the C-Test compared with the cloze test. By deleting every other word, a larger portion and more varied words (parts of speech) are deleted in a C-Test resulting in a more representative sample of language compared to cloze test. Furthermore, a C-Test

is composed of at least four passages, and thus, more text types and text forms can be included, and the chances that some examinees are favored or disadvantaged because of text familiarity are extremely reduced.

C-Test is based on the reduced redundancy principle (RRP; Spolsky, 1971). The RRP basically states that languages contain redundant elements, i.e., elements that get repeated constantly. Redundancy exits in the language to protect it against noise and misunderstanding. When we write and speak, we have a single message, but we repeat the message in different forms and with different words to ascertain that the message is clearly communicated. The RRP was used by Spolsky (1969) and Spolskey et al. (1968) to develop the noise test and later was used to account for other tests including the cloze test and the C-Test. Spolsky argued that since language contains redundancy, a native speaker of a language or a proficient nonnative user of the language should be able to understand it when some parts are deleted. C-Tests have been used in several testing projects as a measure of language ability.

C-Test is used as anchor items in the TestDaf (Eckes & Grotjhan, 2006) which is a standardized test of German as a foreign language for those who want to study in German universities. C-Test is also used as a placement test for learners of German as a foreign language who want to assess their German language knowledge. Numerous researchers have used the C-Test as an overall language proficiency test in studies in second language acquisition (SLA) research (Nadri et al., 2019). In SLA studies, researchers usually need to control for participants' language ability or to measure its impact on other variables of their interest. In other studies, they simply need to know correlates of second language proficiency. In such studies, where administration of a complete language proficiency test like the TOEFL (Test of English as a Foreign Language) or IELTS (International English Language Testing System) is time-consuming or expensive, C-Test is a quick, economical, and precise overall test of second language ability (Norris, 2018).

Over the past four decades, since the introduction of C-Test, numerous studies have shown that C-Test is a valid and reliable measure of first and second language ability (Grotjahn & Drackert, 2020a, 2020b; Motallebzadeh, 2023). Correlational studies have all shown that C-Test highly correlates with other tests of language skills and components even with listening and speaking (Sigott, 2004). Factorial analyses have shown that C-Tests load on a general factor of language proficiency along with other language tests (Grotjahn, 1992; Raatz, 1984; Rasoli, 2021). Other validation methods, like verbal protocol analyses and studies of sensitivity to learning, have indicated that C-Test is a valid and reliable measure of second and first language competence (Baur & Meder, 1994; Borgards & Raatz, 2002; Bolten, 1992; Coleman, 1994; Stemmer, 1991, 1992).

From the early days of the introduction of C-Tests, they have been analyzed with different item response theory (IRT) models (Alpizar et al., 2023; Arras et al., 2002; Baghaei, 2010; Baghaei & Christensen, 2023; Baghaei & Grotjahn, 2014a, 2014b; Eckes & Grotjahn, 2006; Forthmann et al., 2020; Grotjahn, 1992; Rattz, 1984). IRT is a set of psychometric models which define a relationship between a latent trait (like language proficiency) and performance on the items. IRT models rest on three assumptions of unidimensionality, local independence, and monotonicity (Hambleton et al., 1991). If an IRT model fits the data, then persons and items can be located on a unidimensional latent variable with interval scale properties, and the test is measuring a single

latent trait, i.e., unidimensionality holds, and thus, extraneous variables have not contaminated the test, and examinees can be placed on an interval scale. Besides, if an IRT model fits, the raw total score can be used for placing examinees on an ordinal scale (Sijtsma, 1984, 1998).

Mokken scale analysis

Mokken models

Mokken scale analysis (MSA), named after the Dutch mathematician and political scientist Robert J. Mokken, is a set of two nonparametric IRT (NIRT) models for scale construction and a procedure for assessing unidimensionality using the Automatic Item Selection Procedure (AISP, Baghaei, 2021; Tabatabaee-Yazdi et al., 2021). The two NIRT models included under the MSA are the monotone homogeneity model (MHM) and the double monotonicity model (DMM; Mokken, 1971). MHM has the three assumptions of unidimensionality, local independence, and monotonicity. The DMM model contains these three assumptions plus another assumption referred to as nonintersecting item response functions or invariant item ordering (IIO; Sijtsma & Molenaar, 2002).

The first assumption, unidimensionality, suggests that all items are measuring a common underlying trait, denoted as θ (Straat et al., 2013). In the framework of IRT, it is assumed that a single dominant latent trait governs responses to items within a scale, which is referred to as unidimensionality (Hulin et al., 1983). The second assumption is local independence. It implies that an individual's response to one item is entirely unrelated to their responses to any other item. The third assumption, known as the monotonicity of the item response functions (IRF), posits that IRFs are always monotonically increasing functions of the latent trait θ. In simpler terms, this means that as the latent trait level (θ) increases, the probability of an individual providing a correct response on an item also increases. Additionally, the non-intersection assumption, which includes the concept of IIO for dichotomous data (Sijtsma et al., 2011), states that the IRFs of items should not intersect. When the IIO assumption holds true for a set of items, it signifies that the items can be arranged in a hierarchical order from the easiest to the most difficult.

Mokken scale analysis is considered a nonparametric IRT model because it contains no parametric function to define the relationship between the latent trait and item responses. Consequently, unlike parametric IRT models, no person or item parameters can be estimated from the model. In parametric IRT models, such as the Rasch model or the two-parameter logistic model, specific functional forms are assumed for the relationship between item responses and the underlying latent trait. The MHM enables the arrangement of individuals along the latent trait using the sum of their item scores. On the other hand, the DMM not only permits the ordering of individuals based on the sum of their scores but also ensures an invariant ordering of items in terms of difficulty (proportion correct), known as IIO. The IIO property is essential for establishing hierarchical scales, as it ensures that the order of item difficulties remain consistent across all respondents, regardless of their specific values on the latent trait, as highlighted by Sijtsma and Junker (1996).

Scalability coefficients

MSA relies on three scalability coefficients as model fit values, namely, the item-pair scalability coefficient (H_{ij}), the item scalability coefficient (H_i), and the total scale scalability coefficient (H). These coefficients play a crucial role in assessing the quality and properties of a scale. The coefficient of scalability for a pair of items is calculated as the ratio of the covariance between the two items to their maximum obtainable covariance, based on the marginal distribution of the two items. Essentially, H_{ij} measures the internal consistency of each pair of items. H_i represents the ratio of the sum of all pairwise covariances involving item i to the sum of all pairwise maximum covariances related to that specific item. The H_i statistic assesses the scalability of an individual item within the context of the entire set of items. And finally, the homogeneity index for an entire scale H is determined by the ratio of the sum of all pairwise covariances among items to the sum of all pairwise maximum covariances or, alternatively, as the ratio of the sum of all observed errors to the expected errors. The H index provides insight into the internal consistency of the entire scale.

MSA for polytomous items

Mokken scaling has also been extended to analyze polytomous items (Hemker & Sijtsma, 1995; Sijtsma et al., 1990). The underlying principles remain consistent, but the analysis goes beyond just item characteristic curves (ICCs) and involves examining responses at each level within the items, such as the response options on a Likert-type scale. The resulting relationship between these responses and the latent trait score can be represented using item step response functions (ISRFs). ISRFs represent the responses at each step or level of the scale. For instance, in a Likert scale with five response categories, there are four steps between these categories, resulting in four ISRFs. ISRFs play a central role in the analysis of polytomous items using Mokken scaling. The procedure for determining if a set of polytomous items forms a Mokken scale follows a similar procedure to that used for dichotomous items. Similar diagnostic tools also exist to assess whether MHM and DMM hold for polytomous items.

Assessing unidimensionality

Mokken (1971) introduced an automated item selection procedure (AISP) designed to choose multiple items measuring the same trait. AISP is used for automatically searching and identifying unidimensional scales (sets of items) from a larger item pool. AISP begins by selecting the two items that have the largest H_{ij} coefficient (greater than a prespecified cutoff value). Next, items that have H_i values greater than the prespecified cutoff values with the already selected items are added one by one until no items remain that have H_i values larger than the cutoff value. Then, the algorithm repeats this procedure to form a second scale from the remaining items and so on. The algorithm stops when no more items meet these criteria. Items which are not selected are referred to as unscalable.

The present study

As explained before, researchers have mostly used the family of Rasch models (Rasch, 1960/1980) for the analysis of C-Tests. The Rasch model is a parametric IRT (PIRT) model with very strict assumptions. PIRT models, in general, impose a certain mathematical shape for the relationship between the latent trait and the item responses. That

is, the relationship between the latent trait and item response should be of logistic shape (i.e., S-shaped based on the logistic function that is employed to define the relationship between the items responses and the latent trait). If this requirement is not satisfied, the item is rejected as a misfit. The relation between the latent variable and the probability of getting an item right or endorsing a response option can be characterized by a monotonically increasing function (Rajlic, 2020). This function is known as IRF, graphically shown by an ICC. PIRT models are very strict as the IRFs should follow the functional shape that the model imposes.

Mokken (1971) states that PIRT models are appropriate in contexts where the underlying trait which causes the response is deeply understood and known. In contexts where the latent variable is not known, such as affective variables, NIRT models work better. Although it is generally argued and has empirically been shown that the C-Test is an overall measure of first and second language proficiency, several researchers have failed to demonstrate what construct(s) exactly underlies the C-Test (Sigott, 2004). Some researchers have arrived at mixed results concerning whether the C-Test is a micro-level test or a macro-level test (Stemmer, 1991, 1992). Some argue that it is a vocabulary test, while others state that it tests beyond vocabulary and taps deeper linguistic knowledge (Sigott, 2004). Sigott (2004) proposed the fluid construct phenomenon and stated that the C-Test construct changes with person ability and test difficulty. That is, the C-Test construct changes for different test takers depending on the proficiency level. These studies suggest that NIRT models should be a better choice for analyzing C-Tests as the construct underlying the C-Test is not clearly known. Furthermore, Scheiblechner (1999) states that there is little evidence that psychological and mental variables behave according to the logistic function (which is imposed by the PIRT models). NIRT models do not impose a certain functional shape for the ICC and thus are more flexible. The purpose of this study is to show that NIRT models are suitable for C-Tests. The secondary goal of the study is to provide additional validity evidence for C-Test using MSA (Mokken, 1971) as a NIRT model. Specifically, the research questions are addressed:

1. Are C-Tests unidimensional based on the Automatic Item Selection Procedure (AISP) of the MSA?
2. Do C-Tests satisfy the monotonicity assumption?
3. Do C-Tests satisfy the double monotonicity assumption?

Methodology

Participants

For the purpose of this study, 271 (179 female and 92 male) undergraduate university students of English at universities in Khorasan Razavi, Iran, were recruited. Students of TEFL (teaching English as a foreign language), translation, and literature were selected. The age range of the participants was between 18 and 35 ($M = 22.73$, $SD = 3.59$). All participants gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of the Islamic Azad University, Torbat-e Heydarieh Branch.

Instrument

A C-Test battery was developed by the researchers for this study. The C-Test contained six short independent passages. The texts were selected from graded reading comprehension passages available from the British Council website. The reading comprehension exercises on this website are graded and categorized under six levels of A1, A2, B1, B2, C1, and C2. One text was selected from each level. By selecting texts from all levels of the Common European Framework (CEF), we made sure that the C-Test targets all the students at all levels.

The six passages were ordered from the easiest to the hardest based on the British Council website grading. The rule of two (Raatz & Klein-Braley, 2002) was applied to convert the passages into C-Tests. That is, the second half of every second word was deleted, leaving the first and the last sentences in each passage intact to provide enough context for text processing. Twenty words were mutilated in each passage, but proper nouns were not damaged. For words with an even number of letters, the larger parts were deleted. A solid line with a fixed length represented the deleted letters in each word. This means that no clue as regards the number of deleted letters was given to the examinees (Grotjahn, 2019).

Procedure

The C-Test was distributed among undergraduate university students of English in different Khorasan Razavi universities. The C-Test was administrated in reading comprehension courses as a section of their mid-term or final examinations. For scoring, exact word scoring with correct spelling was used. According to Grotjahn (2019), this method of scoring yields the most reliable scores. One point was given for each correct reconstruction. The collected data was entered into Excel, and the following analyses were performed:

1. Dimensionality was evaluated using Mokken's (1971) Automatic Item Selection Procedure (AISP).
2. Reliability was assessed using different methods.
3. The monotone homogeneity model (MHM) of Mokken was checked by examining coefficients of scalability and item response functions.
4. The double monotonicity model (DMM) of Mokken was checked by examining the intersection of all pairs of IRFs graphically and statistically.

The *Mokken* package (van der Ark, 2012) in R (R core team) was used for all the analyses.

Results

Dimensionality and reliability

Table 1 shows the descriptive statistics for the six C-Test passages. To avoid confusion and make full use of the mokken package, the C-Test passages were rescored to have seven categories instead of 21. That is, every three neighboring categories were merged. AISP with 0.30 as the lower bound of scalability coefficient for scale construction

Table 1 Descriptive statistics

	Range	Mean	SD	Skewness	Kurtosis
Item1	5.00	4.78	1.18	−0.90	0.20
Item2	5.00	5.11	1.16	−1.18	0.48
Item3	5.00	4.47	1.44	−0.79	−0.37
Item4	5.00	4.64	1.35	−0.65	−0.78
Item5	5.00	5.04	1.29	−1.17	0.36
Item6	6.00	4.59	1.62	−0.81	−0.58

Table 2 Item and item pair scalability coefficients and their standard errors

	Item1	Item2	Item3	Item4	Item5	Item6
Item1		0.883 (0.023)	0.830 (0.023)	0.847 (0.021)	0.875 (0.022)	0.872 (0.022)
Item2			0.878 (0.022)	0.899 (0.018)	0.877 (0.021)	0.853 (0.029)
Item3				0.880 (0.018)	0.926 (0.014)	0.894 (0.023)
Item4					0.927 (0.013)	0.904 (0.021)
Item5						0.888 (0.029)
Item6						
H_i (SE)	0.860 (0.016)	0.877 (0.016)	0.883 (0.013)	0.892 (0.011)	0.900 (0.014)	0.884 (0.022)

showed that all the items belong to scale 1 which indicates a unidimensional scale. The *Mokken* package estimates different reliability coefficients. The Mokken reliability, alpha, lambda, and LCRC reliabilities were 0.969, 0.965, 0.967, and 0.964, respectively. These coefficients indicate a highly reliable test.

Scalability coefficients

Table 2 shows item and item pair scalability coefficients H_i and H_{ij} and their standard errors (in brackets). As Table 1 shows, all scalability coefficients H_i and H_{ij} are positive and above 0.30. The smallest coefficient is 0.83. Item scalability coefficients H_i are in the last row. H_i is the main statistic to evaluate the MHM (Baghaei, 2021). It indicates item discrimination and fit to the MHM. Items with weak discrimination do not contribute to reliable ordering of examinees and should be discarded. The lower bound for accepting an item as fit to the MHM is $H_i > 0.30$ (Mokken, 1971). As Table 1 shows, all H_i values are positive and above 0.30.

Table 3 shows the monotonicity and IIO statistics. As can be seen, none of the items violates the monotonicity assumption. Column "ac" refers to the number of active comparisons, and column "vi" shows the number of violations which is zero here, i.e., no items violate the assumption of monotonicity. The column "sig" shows the number of significant violation, and column "crit" is a general index of fit. High crit values are not good and show that the items are poor (Baghaei, 2021). Table 3 shows that the crit values for all the items are zero, which is the perfect value.

The scalability coefficient H for the whole test and its standard error were 0.88 and 0.013, respectively. According to Mokken (1971), H-values greater than 0.50 indicate a strong scale.

Table 3 Monotonicity and IIO output

Monotonicity					IIO			
	ac	Vi	Sig	Crit	ac	Vi	Sig	Crit
Item1	10	0	0	0	13	1	1	65
Item2	5	0	0	0	13	0	0	0
Item3	8	0	0	0	15	0	0	0
Item4	3	0	0	0	13	1	1	54
Item5	5	0	0	0	14	0	0	0
Item6	14	0	0	0	14	2	2	95

The statistics, H, H_i, and H_{ij}, play a crucial role in constructing and evaluating the MHM. To account for chance variations, the null hypothesis is tested whether H, H_i, or H_{ij} is equal to zero in the population. This is done by computing the 95% confidence intervals around the coefficients, that is, by adding and subtracting the coefficient's value with two times its standard error. If the interval does not include zero, the null hypothesis that the coefficient is equal to zero in the population can be rejected. A Mokken scale is considered to exist when certain conditions are met:

1. Each pair-wise homogeneity coefficient, H_{ij}, should be greater than 0.
2. Each individual item's homogeneity coefficient, H_i, should be greater than 0.30.
3. Additionally, the overall homogeneity coefficient for the entire scale, H, should also exceed 0.30.

When these three conditions are satisfied, it signifies the presence of monotone homogeneity. This can be summarized as follows: (a) The items within the scale collectively reflect a single underlying latent construct, (b) the assumption of local independence holds, and (c) for each item, the higher a respondent is on the latent construct continuum, the higher the likelihood of a positive response ("person ordering is item-free").

Invariant item ordering

Once a scale is identified as monotone homogeneous, it needs further scrutiny for double monotonicity. Double monotonicity entails that the ordering of items should be consistent across different groups of respondents, or in simpler terms, item response functions should not intersect ("item ordering is person-free").

Table 3 also shows the IIO statistics. "ac" indicates the total number of active pairs, "vi" shows the total number of violations, "sig" shows the number of significant violations, and the "crit" value is a weighted sum of the other elements like "H_i" and "#ac." High "crit" values indicate poor items. "crit" is an index meant as an overall index of fit to the property that is being investigated (0 is perfect; higher is worse). Table 3 shows that item 6 has two violations (with items 1 and 4). In other words, the IRF for this item intersects with the IRF of two other items. Since it has the highest number of violations, it is a good candidate to be removed from the test. Removing this item fixes the intersection of other items with this item. That is, after removing item 6, none of the items has a significant violation of IIO. After omitting item 6, the H_T was 0.34.

Figure 1 shows the IRFs for item pairs. It shows that the IRF for item 6 intersects with the IRFs of items 1 and 4. The IRFs of other items do not intersect. Molenaar and Sijtsma (2000) introduced the coefficient H^T to assess whether curves intersect. H^T is a measure of IIO and stands for scalability coefficient H computed for the transposed data set, that is, a data set in which the position of columns (items) is reversed with the position of rows (persons). If IIO holds, $0.30 < H^T \leq 0.40$ indicates a weak ordering, $0.40 < H^T \leq 0.50$ shows a moderate ordering, and $H^T > 0.50$ shows a strong ordering (Ligtvoet et al., 2010). Under the conditions that there are no (serious) violations of manifest invariant item ordering (MIIO), H^T is a measure of how well the items can be ordered invariantly. Findings showed that H^T value after deleting item 6 was 0.34 which indicates a weak ordering (Ligtvoet et al., 2010). H^T, in fact, provides information about the spread of the IRFs. The farther the IRFs, the higher the H^T is. That is, when IRFs are farther from each other, there is more confidence in IIO.

Discussion

This study explained how the nonparametric monotone homogeneity model and the double monotonicity model contribute to the construction of scales for the measurement of language proficiency. The MHM is more general than the parametric IRT models (Hemker et al., 1997), such as the rating scale model (Andrich, 1978) and the partial credit model (Masters, 1982) which are mostly used for the analysis of C-Tests. Hemker et al. (1997) showed that all parametric IRT models for polytomous items are special cases of the nonparametric MHM. Therefore, any test that satisfies the requirements of a parametric IRT model for polytomous items also satisfies the requirements of the nonparametric MHM. Since the MHM is a more general and a more flexible model compared to its parametric counterparts, its application leads to the retainment of more items and, thus, longer scales. Furthermore, since the total score and the estimated latent trait theta have the same rank order information, the nonparametric MHM is highly applicable for person measurement.

In the C-Test context, we often know very little about the psychometric properties of newly developed batteries. With a typical nonparametric analysis, test developers can examine the dimensionality of the data and evaluate model assumptions such as monotonicity and study the shapes of the IRFs in order to learn more about the (mal-)functioning of individual passages (Baghaei & Effatpamah, 2024; Effatpanah & Baghaei, 2023, 2022). Applying this methodology, researchers can construct scales on which groups can be compared and changes monitored without making unnecessary restrictive assumptions about the behavior of the data (i.e., adherence to a functional form). One reason for misfit of items to parametric IRT models is that the empirical IRFs, although being monotone, do not follow the logistic shape required by many parametric IRT models. Nonparametric MHM and DMM relax the assumption of a functional form for the IRFs and are, therefore, more flexible and less restrictive. However, note that this flexibility comes at a price: nonparametric IRT models do not allow the construction of interval scales, and person and item scores are at ordinal scale level.

When the MHM fits the data, the fit of a parametric model such as the rating scale model or the partial credit model may be investigated. However, if one pursues a parametric IRT model from the outset, misfit may be a good reason to shift to a

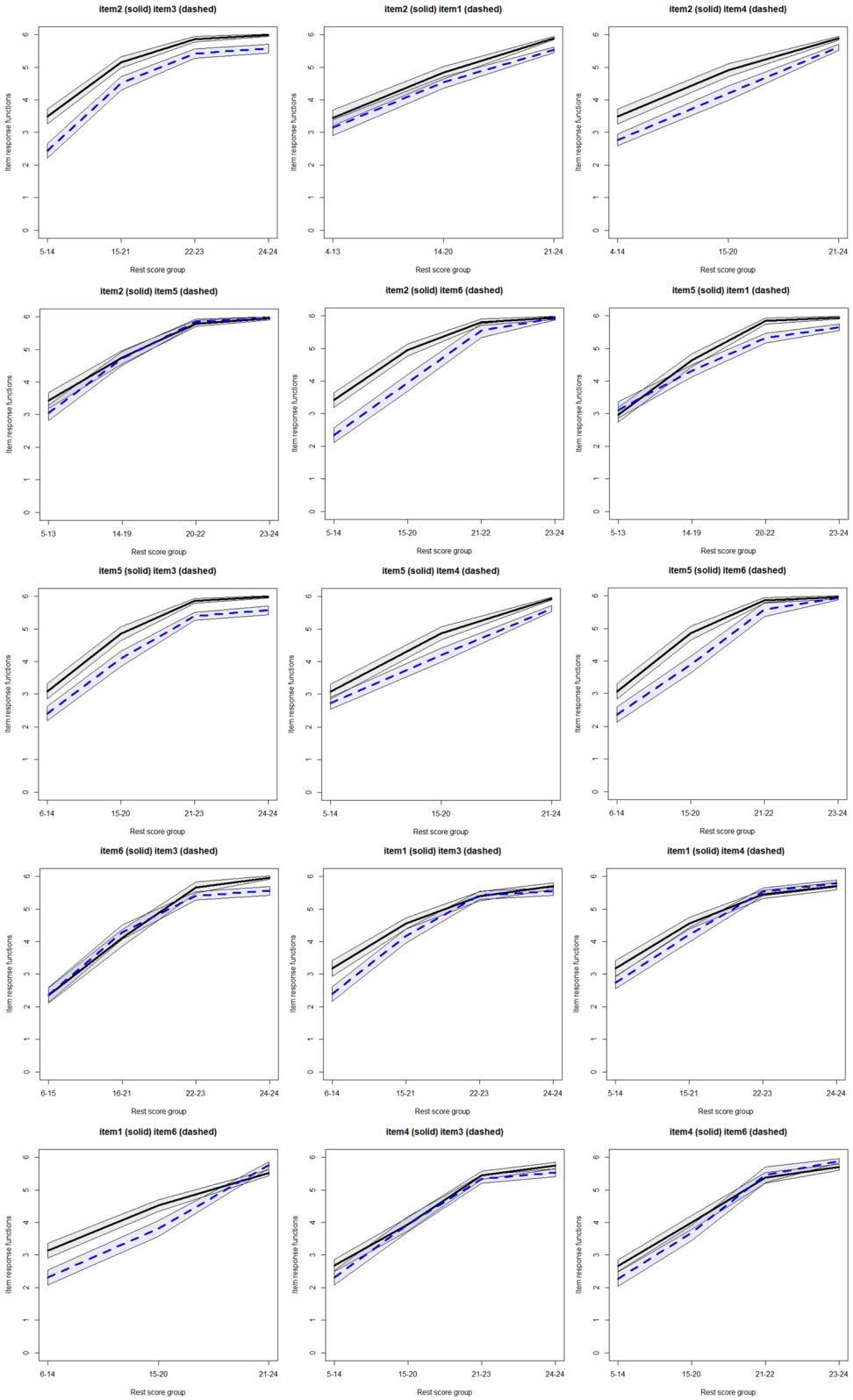

Fig. 1 IRFs for pairs of items to check non-intersection

nonparametric IRT model. An NIRT model still allows the researchers to have an ordinal scale. However, if computer adaptive testing (CAT) is needed to be implemented, a parametric model should be used provided that the model fits the data well.

Overall, the C-Test evaluated in this study showed robust psychometric qualities. More specifically, we found that it has satisfactory monotonicity, scalability, invariant item ordering (with only one significant violation), and reliability. This is an overall satisfactory set of results, which would lead us to encourage the use of C-Tests in language education. Evidence for monotonicity and invariant item ordering along with unidimensionality support the fit of the double monotonicity model.

The fit of the DMM presents several advantages for the use of C-Test in practice (Ligtvoet et al., 2010). First, the monotonicity of item responses suggests that, even if the parametric RSM and PCM do not fit, there is support for the C-Test sum scores to order persons based on their ability. Second, the passages in this C-Test battery were conceptualized to have a hierarchical structure, i.e., they were selected and arranged in order of difficulty based on some theoretical criteria. Evidence for invariant item ordering supports such a hypothetical structure for the test. Test developers and practitioners generally assume, that, because an item *i* has a higher success rate than another item *j*, then item *i* is necessarily easier than *j* for all examinees along the entire range of the scale, and they often use a test as though this assumption was true, without empirically testing it (Ligtvoet et al., 2010). The current study provides evidence that it is empirically justified to make such interpretations from the current C-Test data.

Our findings showed that item 6 had two significant violations of IIO. While usually the recommendation under such circumstances is to remove the intersecting item, we would recommend keeping this item. This is because the IRFs suggest that this items' response function is monotone, and its intersections with the item response functions of items 1 and 4 are very little, as the confidence intervals (the shaded area around the IRFs) overlap for most ability levels (Myszkowski, 2020). The current study suggests that future researchers applying this C-Test use the full instrument, even though they may question and study their own dataset to decide on whether to use item 6 in the scoring or not. Alternatively, the item may be used as a training item at the beginning of the test but is not included in scoring.

Limitations and future directions

A possible reason for the violation of IIO could be the close difficulty of the items. As Table 1 shows, the item means are very close to each other. Consequently, the items do not have much leeway to maneuver (Ligtvoet et al., 2010). This suggests that the instrument can further be improved by selecting passages and constructing C-Test items that differ more in difficulty. This might result in a C-Test whose items can invariably be ordered.

This study has some limitations that stem from the MSA framework. MSA does not provide a way to study or recover information from the response options or categories like other approaches—such as the PCM (Masters, 1982). MSA does graphically provide item step response functions, but because ISRFs in MSA are increasing by definition, it is not possible to evaluate the (mal-)functioning of the response categories. This

is a limiting factor in this context, as previous studies with Likert-type items suggest that some response items may not function well and need to be merged with adjacent categories (Bond et al., 2020). Another limitation of MSA is that it does not provide a way to investigate conditional reliability, and therefore does not allow to monitor if an instrument provides reliable ability scores across a wide range of the ability scale. This is particularly a problem as it implies that the instrument is reliable across the entire range of abilities that are measured. Finally, other advanced uses of PIRT models, such as computer-adaptive testing and test equating, are impossible with Mokken scaling (Meijer et al., 1990).

The current study demonstrates how Mokken scale analysis can provide insightful information about a test form which has already been studied for decades with multiple modern and classical methods (see Grotjahn & Drackert, 2020a, 2020b). We suggest that future studies investigate the psychometric qualities of C-Tests using other nonparametric IRT models—such as, the spline IRT models (Winsberg et al., 1984) and Kernel Smoothing IRT (Ramsay, 1991)—to better understand its functioning.

Conclusion

This study provided valuable insights into the validity of the C-Test format using the MHM and the DMM of the Mokken scale analysis. The main findings indicate that the C-Test analyzed in this study exhibits strong evidence of unidimensionality which is an important aspect of validity (Messick, 1989). Furthermore, all the items in the C-Test conformed to the monotonicity principle of the MHM. The monotonicity principle implies that as a person's proficiency in the construct being measured increases, their likelihood of correctly answering each item also increases. This indicates that there is a causal relationship between the construct and the test scores which is an important aspect of validity in the instrument-based account of validity (Baghaei, 2021; Borsboom et al., 2004). The fit of the MHM implies that persons can be ordered on the latent construct with their C-Test sum scores, and their order is the same regardless of the subset of items that is used. Therefore, the C-Test can be used for placing learners into appropriate language programs, tracking their progress, and making informed decisions regarding their language education.

The fit of the DMM is evidence that the difficulty of the C-Test items remains the same across the ability scale. The implication is that C-Test is fair, invariant, and unbiased. This implies that educators and language teachers can design instructional materials that align with different proficiency levels using C-Tests. A valid and reliable C-Test can be a time-efficient tool for assessing foreign language proficiency, particularly in settings where administering longer, more comprehensive exams may not be practical. This study demonstrates that the C-Test format is a valid and reliable tool for measuring foreign language proficiency. It can support more effective language learning, fair and consistent evaluation, and contribute to the improvement of foreign language education programs and assessments.

Abbreviations

SLA	Second language acquisition
TOEFL	Test of English as a Foreign Language
ILETS	International English Language System

IRT	Item response theory
NIRT	Nonparametric item response theory
PIRT	Parametric item response theory
TEFL	Teaching English as a Foreign Language
CEF	Common European Framework
ICC	Item characteristic curves
MSA	Mokken scale analysis
MHM	Monotone homogeneity model
DMM	Double monotonicity model
AISP	Automatic item selection procedure
LCRC	Latent class reliability coefficient
IIO	Invariant item ordering
IRF	Item response functions
MIIO	Manifest invariant item ordering

Acknowledgements
The authors would like to thank the participants of the study as well as the instructors who collaborated in the data collection process.

Authors' contributions
RS collected the data, performed the analyses, and wrote the first draft of the manuscript. PB conceived, designed, and closely monitored the analyses. HK and HA read the manuscript and contributed to its consistency and coherence.

Funding
The authors did not receive any funding for this research.

Availability of data and materials
The data associated with this study will be available upon request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 29 December 2023 Accepted: 25 February 2024
Published online: 28 March 2024

References

Alderson JC, (1983). The cloze procedure and proficiency in English as a foreign language. In J. W. Jr. Oller (Ed.), Issues in language testing research (pp. 205–217). Newbury House. https://doi.org/10.2307/3586211

Alpizar, D., Li, T., Norris, J. M., & Gu, L. (2023). Psychometric approaches to analyzing C-tests. *Language Testing, 40*(1), 107–132. https://doi.org/10.1177/02655322211062138

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573. https://doi.org/10.1007/bf02293814

Arras, U., Eckes, T., & Grotjahn, R. (2002). C-Tests im Rahmen des Test Deutsch als Fremdsprache (TestDaF): Erste Forschungsergebnisse. In R. Grotjahn (Ed.), *Der C-Test: Theoretische grundlagen und praktische Anwendungen* (Vol. 4, pp. 175–209). Bochum: AKS-Verlag.

Baghaei, P. (2010). An investigation of the invariance of Rasch item and person measures in a C-Test. In R. Grotjahn (Ed.), *Der C-Test: Beiträge aus der aktuellen Forschung/ The C-Test: Contributions from CurrentResearch* (pp. 100–112). Frankfurt/M.: Lang.

Baghaei, P. (2021). *Mokken scale analysis in language assessment*. Münster: Waxmann.

Baghaei, P., & Christensen, K. B. (2023). Modelling local item dependence in C-tests with the loglinear Rasch model. *Language Testing, 40*(3), 820–827. https://doi.org/10.1177/02655322231155109

Baghaei, P., & Effatpanah, F. (2024). Nonparametric kernel smoothing item response theory analysis of Likertitems. *Psych, 6*(1), 236–260. https://doi.org/10.3390/psych6010015

Baghaei, P., & Grotjahn, R. (2014a). Establishing the construct validity of conversational C-Tests using amultidimensional Item Response Model. *Psychological Test and Assessment Modeling, 56*, 60–82.

Baghaei, P., & Grotjahn, R. (2014b). The validity of C-Tests as measures of academic and everyday language proficiency: A multidimensional item response modeling study. In R. Grotjahn (Ed.). *Der C-Test: Aktuelle Tendenzen/The C-Test: Current trends* (pp. 163-171.). Frankfurt/M.: Lang.

Baur, R. S., & Meder, G. (1994). C-Tests zur ermittlung der globalen sprachfähigkeit im Deutschen und in der muttersprache bei ausländischen schülern in der bundesrepublik Deutschland. In R. Grotjahn (Ed.), *Der C-Test: Theoretische grundlagen und praktische anwendungen* (Vol. 2, pp. 151–178). Bochum: Brockmeyer.

Bolten, J. (1992). Wie schwierig ist ein C-Test? Erfahrungen mit dem C-Test als einstufungstest in hochschulkursen Deutsch als fremdsprache. In R. Grotjahn (Ed.), *Der C-Test. Theoretische grundlagen und praktische anwendungen* (Vol. 1, pp. 193–203). Bochum: Brockmeyer.

Bond, T. G., Yan, Z., & Heene, M. (2020). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th Ed.). New York: Routledge.

Borgards, S., & Raatz, U. (2002). Sind C-Tests trainierbar? In R. Grotjahn (Ed.), *D er C-Test: TheoretischeGrundlagen und praktische Anwen-dungen* (Vol. 4, pp. 157–174). Bochum: AKS-Verlag.
Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061
Coleman, J. A. (1994). Degrees of proficiency: assessing the progress and achievement of university language learners. *French Studies Bulletin, 50*, 11–16.
Eckes, T., & Grotjahn, R. (2006). A closer look at the construct validity of C-tests. *Language Testing, 23*, 290–325. https://doi.org/10.1093/frebul/15.50.11
Effatpanah, F., & Baghaei, P. (2022). Exploring rater quality in rater-mediated assessment using the nonparametricitem characteristic curve estimation. *Psychological Test and Assessment Modeling, 64*(3), 216–252.
Effatpanah, F., & Baghaei, P. (2023). Kernel smoothing item response theory in R: A didactic.P ractical Assessment, Research, and Evaluation, 28, Article 7. https://doi.org/10.7275/pare.1261
Forthmann, B., Grotjahn, R., Doebler, P., & Baghaei, P. (2020). A comparison of different item response theory models for scaling speeded C-tests. *Journal of Psychoeducational Assessment, 38*, 692–705. https://doi.org/10.1177/0734282919889262
Grotjahn, R., & Drackert, A. (2020). *The electronic C-test bibliography: Version October 2020*. Available at: http://www.c-test.de & https://www.ruhr-uni-bochum.de/sprachetesten/index.html.de
Grotjahn, R. (1992). Der C-Test: Einleitende Bemerkungen. In R. Grotjahn (Ed.), *Der C-Test: Theoretische grundlagen und praktische anwendungen* (Vol. 1, pp. 1–18). Bochum: Brockmeyer.
Grotjahn, R. (2019). C-Tests. In S. Jeuk & J. Settinieri (Eds.), *Sprachdiagnostik Deutsch als zweitsprache: Ein handbuch* (pp. 579–603). De Gruyter Mouton.
Grotjahn, R., & Drackert, A. (2020). *The electronic C-test bibliography: Version October 2020*. Available at http://www.c-test.de
Hambleton, R., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory*. Sage.
Hemker, B. T., & Sijtsma, K. (1995). Selection of unidimensional scales from a multidimensional item bank in the polytmous Mokken IRT model. *Applied Psychological Measurement, 19*, 337–352.
Hemker, B. T., Sijtsma, K., Molenaar, I. W., & Junker, B. W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika, 62*(3), 331–347. https://doi.org/10.1007/bf02294555
Hulin, L. H., Drasgow, Y., & Parsons, C. K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irvin.
Klein-Braley, C. (1985). A cloze-up on the C-test: a study in the construct validation of authentic tests. *Language Testing, 2*(1), 76–104. https://doi.org/10.1177/026553228500200108
Ligtvoet, R., van der Ark, L. A., te Marvelde, J. M., & Sijtsma, K. (2010). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement, 70*, 578–595. https://doi.org/10.1177/0013164409355697
Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174. https://doi.org/10.1007/BF02296272
Meijer, R. R., Sijtsma, K., & Smid, N. G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. *Applied Psychological Measurement, 14*(3), 283–298. https://doi.org/10.1177/014662169001400306
Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13 103). New York, NY: American Council on Education and Macmillan.
Mokken, R. J. (1971). *A theory and procedure of scale analysis*. De Gruyter. https://doi.org/10.1515/9783110813203
Molenaar, W., & Sijtsma, K. (2000). *MSP5 for Windows user's manual*. Groningen: Iec ProGAMMA.
Motallebzadeh, Z. (2023). A comparison of different methods for investigating the reliability of C-tests. *Educational Methods & Practice, 1*, 1.
Myszkowski, N. (2020). A Mokken scale analysis of the last series of the standard progressive matrices (SPM-LS). *Journal of Intelligence, 8*(2), 22. https://doi.org/10.3390/jintelligence8020022
Nadri, M., Baghaei, P., & Zohoorian, Z. (2019). The contribution of cognitive abilities and general language proficiency to explaining listening comprehension in English as a foreign language. *Cogent Education, 6*(1), 156710. https://doi.org/10.1080/2331186X.2019.1567010
Norris, J. M. (2018). *Developing C-tests for estimating proficiency in foreign language research*. Frankfurt am Main: Peter Lang.
Raatz, U. (1984). The factorial validity of C-tests. In T. Culhane, C. Klein-Braley, & D. K. Stevenson (Eds.), *Practice and problems in language testing 7. Proceedings of the seventh international language testing symposium of the IUS, Colchester, October 1983* (pp. 124–139). Colchester: University of Essex, Department of Language and Linguistics.
Raatz, U. & Klein-Braley, C. (1982). The C-test – A modification of the cloze procedure. In T. Culhane, C. Klein-Braley & D. K. Stevenson (Eds.), *Practice and problems in language testing IV. Proceedings of the Fourth International Language Testing Symposium of the Interuniversitäre Sprachtestgruppe* (pp. 113–138). Colchester: University of Essex, Dept. of Language and Linguistics.
Raatz, U., & Klein-Braley, C. (2002). Introduction to language testing and to C-tests. In J. A. Coleman, R. Grotjahn, & U. Raatz (Eds.), *University language testing and the C-test* (pp. 75–91). AKS-Verlag.
Rajlic, G. (2020). Visualizing items and measures: An overview and demonstration of the Kernel smoothing item response theory technique. *The Quantitative Methods for Psychology, 16*(4), 363–375. https://doi.org/10.20982/tqmp.16.4.p363
Ramsay, J. O. (1991). Kernel smoothing approaches to non-parametric item characteristic curve estimation. *Psychometrika, 56*, 611–630.
Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research, 1960. (Expanded edition, Chicago: The university of Chicago Press, 1980).
Rasoli, M. K. (2021). Validation of C-test among Afghan students of English as a foreign language. *International Journal of Language Testing, 11*(2), 109–121.
Scheiblechner, H. (1999). Additive conjoint isotonic probabilistic models. *Psychometrika, 64*, 295–316. https://doi.org/10.1007/BF02294297
Sigott, G. (2004). *Towards identifying the C-Test construct*. Peter Lang.
Sijtsma, K. (1984). Useful nonparametric scaling: A reply to Jansen. *Psychologische Beiträge, 26*, 423–437.
Sijtsma, K. (1998). Methodology review: Nonparametric IRT approaches to the analysis of dichotomous item scores. *Applied Psychological Measurement, 22*, 3–31. https://doi.org/10.1177/014662169802210001

Sijtsma, K., Debets, P., & Molenaar, I. W. (1990). Mokken scale analysis for polychotomous items: theory, a computer program and an empirical application. *Quality and Quantity, 24*, 173–188. https://doi.org/10.1007/BF00209550
Sijtsma, K., & Junker, B. W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical and Statistical Psychology, 49*, 79–105. https://doi.org/10.1111/j.2044-8317.1996.tb01076.x
Sijtsma, K., Meijer, R. R., & van der Ark, L. A. (2011). Mokken scale analysis as time goes by: an update for scaling practitioners. *Personality and Individual Differences, 50*, 31–37. https://doi.org/10.1016/j.paid.2010.08.016
Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Sage. https://doi.org/10.4135/9781412984676
Spolsky, B. (1969, September 8–12). *Reduced redundancy as a language testing tool* [Conference presentation]. Second International Congress of Applied Linguistics, Cambridge, England. https://eric.ed.gov/?id=ED031702.
Spolsky, B. (1971). Reduced redundancy as a language testing tool. In G. E. Perren & J. L. M. Trim (Eds.), *Applications of linguistics* (pp. 383–390). Cambridge University Press.
Spolsky, B., Bengt, S. M., Sato, E. W., & Aterburn, C. (1968). Preliminary studies in the development of techniques for testing overall second language proficiency. *Language Learning, 18*(3), 79–101. https://doi.org/10.1111/j.1467-1770.1968.tb00224.x
Straat, J. H., van der Ark, L. A., & Sijtsma, K. (2013). Comparing optimization algorithms for item selection in Mokken scale analysis. *Journal of Classification, 30*, 75–99. https://doi.org/10.1007/s00357-013-9122-y
Stemmer, B. (1991). *What's on a C-test taker's mind: Mental processes in C-test taking*. Bochum: Brockmeyer.
Stemmer, B. (1992). An alternative approach to C-test validation. In R. Grotjahn (Ed.), *Der C-Test: Theoretische grundlagen und praktische anwendungen* (Vol. 1, pp. 97–144). Bochum: Brockmeyer.
Tabatabaee-Yazdi, M., Motallebzadeh, K., & Baghaei, P. (2021). A Mokken scale analysis of an English reading comprehension test. *International Journal of Language Testing, 11*(1), 132–143.
vanderArk, L. A. (2012). New developments in Mokken scale analysis in R. *Journal of Statistical Software, 48*, 1–27. https://doi.org/10.18637/jss.v048.i05
Winsberg, S., Thissen, D., & Wainer, H. (1984). Fitting item characteristic curves with spline functions. *ETS Research Report Series, 1984*(2), i–14. https://doi.org/10.1002/j.2330-8516.1984.tb00080.x

Publisher's Note