The Journal of English Language Pedagogy and Practice
Vol. 16, No.33, Autumn and Winter 2023(148-173)

Research Article

Designing an Intercultural Development Inventory to Assess EFL Learners' Intercultural Competence: A Mokken Scale Analysis

Mahboubeh Akbari*[1], Mona Tabatabaee-Yazdi[2]
[1,2]*English Department, Tabaran Institute of Higher Education, Mashhad, Iran*
*Corresponding author: Mahboubehakbari82@gmail.com
(Received: 2023/10/07; Accepted: 2024/05/28)
Online publication: 2024/12/08

Abstract

While the issue of intercultural competence has received considerable attention in second language learning and teaching over recent decades, the lack of a practical, valid, and dependable instrument to cross-culturally measure intercultural competence is strongly felt. Hence, the current study aimed at developing and validating an inventory to examine EFL learners' intercultural competence within the context of Iran. To this end, following a qualitative-quantitative descriptive research design, an instrument consisting of items adapted from intercultural competence-based survey instruments was generated. The study recruited 200 Iranian EFL learners to fill out the inventory. Furthermore, nine Iranian university professors in Applied Linguistics received an earlier draft of the Intercultural Competence Inventory (ICI) after it was prepared. The experts were requested to provide feedback on the developed inventory' content and face validity. After expert/content validation, the construct validity of the inventory was checked using R software running Mokken Scale Analysis. The analyses aimed to determine the structure of the inventory and verify the effectiveness of a five-point Likert response scale. The results concerning item attributes and the quality of the response scale provided confirmation of the scale's internal validity. Hopefully, the presence of this inventory in the Iranian context, characterized by strong internal consistency, internal validity, and construct validity, can contribute valuable insights to the study of intercultural competence among language learners.

***Keywords*:** intercultural competence, Mokken Scale Analysis, reliability, validation

Introduction

Briones and Ramos (2021) defined culture as the focus on the way a social group represents itself through works of art, literature, social institutions, and the process of their production and preservation throughout history. Gray et al. (2019) also stated that culture is a group of shared ideas, values, formation, and assumptions about life that are consciously or unconsciously accepted as *correct* by people in the society. Recently, Kim (2019) noted that foreign language teaching and learning have taken an intercultural turn. In this regard, it is now widely acknowledged that language teaching should not only focus on linguistic and communicative competence but also intercultural competence. Regarding intercultural competence, it is essential for students to recognize that culture distinguishes one community from another. An individual's background shapes their identity, influences their self-perception, their perceptions of others, and their interpretation of the world's realities (Kim, 2019). The relation between language and culture and students' perception of target culture integration is one of the concerns of linguistic researchers (Chao, 2013). That is why linguistics scholars have argued cultural competence is as important as linguistic competence, especially for foreign language learners (Tran & Duong, 2018). Similarly, Seelye (1993) notes that the study of language cannot be detached from the study of culture; neither language nor culture can be taught separately. As such, cultural subjects must be included as a part of any language course so that learners will be able to handle their cultural misunderstandings, monitor the foreign culture, and reflect on their own (Zhang & Zhou, 2019).

Given that culture deals with human life, it is an area of interest for researchers in several domains such as anthropology, sociology, and education. According to Merrouche and Adberrahim (2006), anthropologists view culture as a unifying force while Brooks (1968) defines it as the sum of all the learned and shared elements that characterize a societal group. Culture, an inherent aspect of human society, encompasses various practices and forms of information and plays a significant role in the interpretation of meaning. This results in people from the same cultural group having unique expressions that distinguish them from others. In other words, everyone reflects on own special thoughts and culture (Schwieter et al., 2021). In a general sense, culture could also be vastly defined as sets of actions and behaviors that affect the life of members of society, common beliefs, acceptable traditions, and ways of life which all can form and transform the characters of the society's members (Geertz, 2000). In this regard, Hoftsede

(2003) defines culture as a set of acquired and transmitted ways of thinking, feeling, and reacting that are primarily communicated through symbols, and that constitute the unique accomplishments of human groups. Hence, the essential core of culture lies in traditional ideas and especially their attached values. Samovar and Porter (2004, p. 29) define culture as "... culture is both teacher and textbook" since when someone practices the language outside the native context, a textbook has to make learners more aware of that target culture. Besides, human beings are ultra-social species who are required to interact with each other to meet their everyday needs via using language. This latter is considered as the primary medium of communication which uses an arbitrary system produced voluntarily and covers both verbal and non-verbal aspects such as sounds, gestures, and written or spoken symbols. Hence, Rangriz and Harati (2017) defined language as "a purely human and non-instinctive method of communicating ideas, emotions, and desires by means of a system of voluntarily produced symbols" (p. 209). Additionally, language cannot be visualized in a vacuum, in this logic; any language has a setting which basically is a society or culture. Therefore, language and culture intricately intertwine and coexist, influencing and shaping one another seamlessly, as elucidated by the findings in the work of Merrouche and Abderrahim (2006). The National Standards for Foreign Language Education Project in 2006 emphasized that mastering the culture of a foreign language is essential for students to effectively learn that language. This underscores the importance of cultural comprehension in attaining a high level of proficiency in a foreign language. Singhal (1997) emphasized that the journey of learning a foreign language cannot be deemed truly fulfilled without the inclusion of cultural education. She highlighted the significance of imparting cultural knowledge to students as an essential component that offers a profound context for learners to proficiently utilize the language. In the same vein, Brown (2020) described this connectedness as: "Language is a part of a culture, and culture is a part of the language; the two are intricately interwoven so that one cannot separate the two without losing the significance of either language or culture" (p. 169).

Therefore, there is no doubt that the interrelationship between language and culture is entrenched and profoundly rooted. On that basis, they are tightly and closely correlated with each other. On the one hand, language is the specific human vehicle by which a culture conveys its beliefs, values, and norms. In this light, language conducts social lives to convey culture and preserve people's cultural ties and serves as "a communication mechanism that embodies expresses and symbolizes cultural reality" (Kramsch, 2014, p.

3). In contrast, language is regarded as the product of culture; it is just one of the various cultural products (Muir, 2007).

The concept of Cultural Competence (CC) comprises culture and competence. The former involves humans' behaviors, ways of communicating, beliefs, and what makes them worth. The latter indicates the required capacity to perform in a prosperous way in the communication process (Cross, 1989). In assigning definitions to CC, the emphasis is mostly on individuals' abilities to be aware of the diversities among the ethnic groups to respect and deal with the differences properly. Accordingly, five major elements were made by Cross (1989) that have contributory roles in making people, institutions, and professionals culturally competent agents. The first element gives an estimation of the diversities that exist by giving respect to others who belong to different cultures. The second one promotes individuals' awareness towards self-assessment which means people should know and value their own culture. Awareness about the dynamics when interacting with others is explained as the third factor, meaning that many things may change because of individuals' interactions. The fourth element focuses on the application and institutionalization of knowledge, and the final element explains diversity adaptation in addition to the context of the served culture.

Moreover, CC is not achieved by obliging people to accept others' cultures and behave in the same way as they do (Kramsch, 2014). Baraja-Rohan (1999) argued that CC is about enabling learners to be aware of cultural diversities, to acquire the ability to notice and accept the differences that exist between people, and the ways to overcome the divergence in a successful manner. Accordingly, being culturally competent refers to the capability to experience and distinguish the variances that may exist between cultures. Thus, intercultural sensitivity is construed as the first reaction that individuals produce whenever they face an intercultural situation as well as predict cultural competence (Altshuler et al., 2003). The analysis and interpretation of the results highlight the importance of language proficiency, cultural background, and prior intercultural experiences as influential elements in promoting intercultural competence. These findings carry implications for educators and policymakers, underscoring the necessity of designing language and intercultural education programs that incorporate these key elements to effectively enhance learners' cross-cultural skills. Likewise, in a study conducted by Schat et al. (2021), they focused on creating and validating an evaluation instrument for assessing intercultural competence in upper-secondary foreign language teaching. By employing both exploratory and

confirmatory factor analysis, the researchers assessed the construct validity of the instrument, utilizing a sample of 164 students in the Netherlands. Although the findings offered confirmation for the anticipated second-order factor arrangement, the indices evaluating the model's fit were not as favorable as those found in an alternative model featuring five first-order factors. Besides, Duisembekova (2021) developed and validated a 34-item instrument to explore English language teaching student teachers' beliefs about intercultural communicative competence in Turkey. The instrument demonstrated excellent reliability and identified four factors: attitudes, knowledge, awareness, and skills. Tabatabaee-Yazdi and Baghaei (2022) evaluated the Persian translation of the Intercultural Intelligence Scale using 203 Iranian EFL teachers. The questionnaire comprised four dimensions and showed good fit with the Rasch model indicating the effective measurement of teachers' intercultural intelligence.

Hence, with a recognized gap in the availability of a trustworthy and effective tool for evaluating students' intercultural skills (Alijanian et al., 2019; Ramos & Briones, 2021; Zhang & Zhou, 2019), there arises a critical need to develop such an instrument to advance research in this area. Consequently, the primary objective of the current study was to create and validate an assessment tool focused on assessing EFL learners' intercultural competence. Therefore, the following research questions have been posed:

Q1: Does the developed inventory on EFL students' intercultural competence have construct validity?

Q2: Does the developed inventory of EFL students' intercultural competence have reliability?

Method

Participants and Settings

In previous studies in the same field, the sample size of participants varied from 133 to 15022. Straat et al. (2014) recommend a sample size of 250 to 500 respondents when the item quality is high and 1250 to 1750 when the item quality is low while Wright and Stone (1979) state that the results are more trustworthy when the dataset is closer to the center of the 100-point distribution. Therefore, this study invited 200 participants to fill out the inventory. They were studying EFL and came from diverse educational backgrounds, including both university and institute settings. They enjoyed diverse English language proficiency levels including pre-intermediate (%11), intermediate (%30.2), high intermediate (%32.2), and advanced

(%26.2). It is worth noting that the participants with elementary English proficiency levels were not included in the study as they would not have been able to answer the questionnaire due to their limited language skills. The participants included both genders, females (%70.5) and males (%29.5), within different age groups (Mean $_{age}$= 29.23; SD= 7.50),

Instruments and Scale Development

The instrument used in this study was developed based on an extensive review of existing literature and validated scales related to the constructs under investigation. Key sources, including (Chao, 2014; Deardorff, 2006; DeJaeghere & Zhang, 2008; Jiao et al., 2020; González-López & Fernández-Montoto, 2018; Griffith et al., 2016; Günçavdi & Polat, 2016; Stemler et al, 2014; Tabatabaee-Yazdi & Baghaei, 2022; Wang et al., 2022), were consulted to ensure that the items accurately reflected the theoretical foundations of behavioral, cognitive, and motivational frameworks (see Appendix A). The instrument was divided into two sections. The first section gathered demographic details of the participants, while the second section consisted of 41 items that measured three distinct constructs: behavioral and skills (items 1-19), cognitive and metacognitive (items 20-36), and motivational (items 37-41). The responses were set on a five-point Likert scale (ranging from strongly disagree to strongly agree). Mokken Scale Analysis was employed to evaluate the construct validity and reliability of the questionnaire. The entire inventory took participants approximately 20 minutes to complete.

Study Design and Data Analyses

This study was qualitative-quantitative exploratory descriptive research. Accordingly, the study recruited 200 Iranian EFL learners to fill out the inventory. Furthermore, 9 Iranian university professors in Applied Linguistics received an earlier draft of the Intercultural Competence Inventory (ICI) after it was designed. The experts were requested to provide feedback on the developed Inventory' content and face validity. In the first phase, in order to examine the most recent conceptions of intercultural competence in various educational fields, a comprehensive literature review was conducted. In order to create the Intercultural Competence Questionnaire (ICQ), a closer look at the existing instruments used to assess students' intercultural competence was conducted. Nine experts in Applied Linguistics received an earlier draft of the ICQ after it was prepared. The experts were asked to provide feedback on the developed questionnaire's language and face validity. In addition, they were requested to provide feedback on the questionnaire's presentation of the construct of learners' intercultural competence and suggest additional items if they believed it had been

misrepresented. For example, according to one of the experts, about 20 items were considered duplicates and were delete. According to this expert, there was a lot of similarity between repeated items, which should have been merged and new items were presented. Another feedback from an expert was that the words used in the items were unfamiliar words for the people who want to answer them. After expert/content validation, the construct validity of the inventory was checked using R software running Mokken Scale Analysis.

Preliminary Analyses

As a preliminary check, descriptive statistics were calculated including the mean, standard deviation (SD), skewness, and kurtosis to describe the data in terms of their distribution. As illustrated in Table 1, items 28 (M = 2.73, SD = 1.113, total score = 543) and 31 (M = 2.61, SD = 0.967, total score = 520) have the lowest mean scores, and items 38 (M = 4.05, SD = 0.955, total score = 805) and 39 (M = 3.91, SD = 1.026, total score = 778) have the highest. The skewness and kurtosis values for all items fall within the acceptable range of -2 to +2, indicating that the data are symmetric in shape. Besides, the degree of reliability for the scale was investigated utilizing the Cronbach alpha coefficient (1951), and the value of 0.91 was obtained, showing strong internal consistency reliability for the scale.

Table 1
Descriptive Statistics for the Scale Data

Items	Mean	Standard Deviation	Skewness	Kurtosis
1	3.47	1.091	-0.517	-0.475
2	3.56	1.003	-0.418	-0.479
3	3.79	0.883	-0.738	0.586
4	3.71	0.972	-0.619	-0.030
5	3.63	1.059	-0.535	-0.305
6	3.51	0.893	-0.583	0.275
7	3.85	0.945	-0.902	0.791
8	3.24	1.026	-0.329	-0.350
9	3.62	1.032	-0.586	-0.246
10	3.24	0.932	-0.223	-0.046
11	3.48	0.979	-0.524	-0.089
12	2.79	1.217	0.293	-0.855
13	3.39	0.998	-0.654	0.117
14	3.34	0.918	-0.257	-0.201
15	3.36	0.937	-0.248	-0.284

16	3.56	1.007	-0.653	0.049
17	3.62	0.955	-0.901	0.695
18	3.48	0.974	-0.597	-0.187
19	3.83	0.954	-0.674	0.222
20	3.22	1.029	-0.248	-0.477
21	3.27	0.968	-0.332	-0.331
22	3.41	0.995	-0.406	-0.195
23	3.34	0.955	-0.452	-0.121
24	3.09	1.016	-0.183	-0.497
25	3.14	1.013	-0.129	-0.703
26	3.30	1.020	-0.114	-0.877
27	3.16	0.995	-0.287	-0.485
28	2.73	1.113	0.265	-0.685
29	3.46	1.018	-0.405	-0.482
30	3.48	0.953	-0.590	-0.062
31	2.61	0.967	0.336	-0.307
32	3.20	1.104	-0.214	-0.701
33	2.99	1.135	-0.241	-0.894
34	3.04	1.037	-0.016	-0.792
35	2.98	1.082	-0.081	-0.580
36	3.21	0.956	-0.295	-0.260
37	3.37	0.970	-0.424	-0.137
38	4.05	0.955	-1.215	1.612
39	3.91	1.026	-0.866	0.345
40	3.65	0.987	-0.402	-0.453
41	3.38	1.148	-0.353	-0.692
Total Score	138.44	19.400	-0.442	0.689

Figure 1 further illustrates the distribution of the responses for each category of the scale. As can be seen, response option 1 (Strongly Disagree) had the lowest percentage (5%), suggesting that a small proportion of respondents have endorsed this category. This response option was followed by response options 4 (Strongly Agree) and 2 (Disagree) with 13 and 16 percent, respectively. However, response option 4 (Agree) had the highest percentage (38%), indicating that a large number of the respondents have endorsed the category, followed by response option 3 (Neutral) with 28 percent. This can be considered as evidence that most of the respondents have a higher intercultural competence.

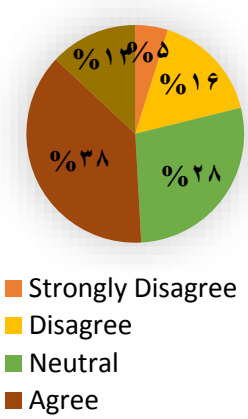

Figure 1. *Distribution of Responses for Each Category of the Scale*

Scalability Coefficients

The data were analyzed using the Mokken package version 3.0.6 (van der Ark et al., 2022) in R (R Core Team, 2013). To assess whether the items of the scale could form a Mokken scale, the scalability coefficients for all items in the scale (H), the scalability coefficient for each individual item in the scale (Hi), and the scalability coefficients for each item-pair (Hij) were first examined. Following the criteria suggested by Mokken (1971), a scale is considered weak if $0.30 \leq H \leq 0.40$, medium if $0.40 \leq H \leq 0.50$, and strong if $H \geq 0.50$; the values of Hij must be greater than zero or non-negative; items should be reviewed or deleted if the coefficient is $Hj < 0.30$, but if $Hj \geq 0.30$, they should be selected to form a Mokken scale. For inter-item pairs, the results of the inter-item scalability coefficients (Hij) showed that the values of certain item pairs were negative and less than 0.30. This indicates a negative relationship between an item and the latent trait being measured. Table 2 presents the results of scalability coefficients for the items and the whole scale along with their standard errors. For items, the scalability coefficients (H_i) ranged from 0.118 to 0.329. Except for item 11, the scalability coefficients for the other items were below 0.30. For the whole scale, the scalability coefficient was 0.223 (SE = 0.024), indicating a weak scale.

Table 2
Item Scalability Coefficients of the Scale Items

Items	Scalability Coefficients	Standard Errors (SE)
1	0.251	0.036
2	0.294	0.033
3	0.262	0.040
4	0.256	0.033
5	0.211	0.035
6	0.274	0.035
7	0.208	0.042
8	0.214	0.037
9	0.242	0.036
10	0.252	0.039
11	0.329	0.031
12	0.118	0.039
13	0.280	0.029
14	0.252	0.034
15	0.262	0.033
16	0.248	0.037
17	0.235	0.034
18	0.254	0.037
19	0.249	0.038
20	0.231	0.036
21	0.246	0.032
22	0.265	0.033
23	0.230	0.036
24	0.248	0.037
25	0.208	0.037
26	0.228	0.037
27	0.196	0.036
28	0.136	0.042
29	0.237	0.035
30	0.258	0.035
31	0.146	0.046

32	0.158	0.038
33	0.144	0.043
34	0.124	0.040
35	0.179	0.043
36	0.143	0.040
37	0.209	0.035
38	0.258	0.031
39	0.268	0.032
40	0.257	0.033
41	0.142	0.037
Scale	0.223	0.024

Automated Item Selection Procedure (AISP)

To assess whether the scale measures a single latent trait, the MSA utilizes an automated item selection procedure (AISP). Unidimensionality is the concept that all items in a scale should measure the same underlying trait. The AISP analysis helps identify a group of scalable items that measure this latent trait and adhere to the monotone homogeneity model. Similar to exploratory factor analysis (EFA), AISP divides the data into subscales that meet MSA criteria, potentially including some unscalable items (Baghaei, 2021). AISP can be used to identify and remove non- or low-discriminating items (Sijtsma & van der Ark, 2017). The results of AISP for the scale are given in Table 3. The value 0.30 on the top shows the lower bound of the scalability coefficient for constructing scales (Sijtsma & van der Ark, 2017). Zero indicates that the item is unscalable, and '1', '2', '3', '4', '5', '6', and '7' show that the item belongs to scales 1, 2, 3, 4, 5, 6, and 7, respectively. As can be seen, from the 41 items, five items were unscalable, nineteen items formed a unidimensional scale, and the rest of the items formed short scales.

Table 3
The Results of the Automated Item Selection Procedure (AISP) for the Scale

Items	**Dimensions from AISP c = 0.30**	**Items**	**Dimensions from AISP c = 0.30**
1	1	**22**	1
2	1	**23**	5
3	1	**24**	5
4	1	**25**	6
5	1	**26**	6

6	1	**27**	7
7	1	**28**	7
8	0	**29**	3
9	1	**30**	3
10	1	**31**	4
11	1	**32**	0
12	0	**33**	4
13	1	**34**	4
14	1	**35**	2
15	2	**36**	2
16	1	**37**	3
17	0	**38**	1
18	1	**39**	1
19	1	**40**	1
20	5	**41**	0
21	5		

Note. c = 0.30 is the cut-off value or lower bound of the scalability coefficient.

Monotonicity

Monotonicity is a crucial assumption in MSA, stating that as individuals' ability level (e.g., θ) increases, their likelihood of providing a correct response or endorsing a higher response option should increase. Table 4 illustrates the analysis of the monotonicity assumption: The second column (#ac') shows the total number of the active pairs of the rest score groups used to test manifest monotonicity; the third column (#vi) displays the total number of violations; the fourth column (#vi/#ac) presents the average number of violations per active pair; columns five (maxvi) and six (sum) represent the maximum violation and sum of all violations, respectively; the seventh column (sum/#ac) demonstrates the average violation per active pair; columns eight (zmax) and nine (#zsig) respectively indicate the maximum test statistic and a number of significant violations; and the last column (crit) shows a weighted sum involving elements like 'item H', '#ac', etc. A high 'crit' value indicates poor items. According to Molenaar and Sijtsma (2000), 'crit' values below 0.40 suggest that items adhere to the monotonicity hypothesis while values above 0.40 indicate a violation of monotonicity.

Table 4
The Results of the Monotonicity Assessment

Items	#ac	#vi	#vi/#ac	maxvi	sum	sum/#ac	zmax	#zsig	crit
1	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
2	3	0	0.00	0.00	0.00	0.0000	0.00	0	0
3	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
4	3	0	0.00	0.00	0.00	0.0000	0.00	0	0
5	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
6	3	0	0.00	0.00	0.00	0.0000	0.00	0	0
7	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
8	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
9	3	0	0.00	0.00	0.00	0.0000	0.00	0	0
10	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
11	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
12	12	1	0.08	0.06	0.06	0.0053	0.67	0	35
13	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
14	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
15	3	0	0.00	0.00	0.00	0.0000	0.00	0	0
16	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
17	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
18	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
19	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
20	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
21	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
22	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
23	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
24	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
25	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
26	9	0	0.00	0.00	0.00	0.0000	0.00	0	0
27	4	0	0.08	0.00	0.00	0.0000	0.00	0	0
28	12	1	0.00	0.03	0.03	0.0026	0.41	0	26
29	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
30	3	0	0.00	0.00	0.00	0.0000	0.00	0	0
31	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
32	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
33	12	1	0.08	0.03	0.03	0.0026	0.37	0	26

34	12	2	0.17	0.11	0.15	0.0126	1.15	0	59
35	12	2	0.17	0.05	0.10	0.0084	0.69	0	44
36	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
37	12	0	0.00	0.00	0.00	0.0000	0.00	0	0
38	3	0	0.00	0.00	0.00	0.0000	0.00	0	0
39	4	0	0.00	0.00	0.00	0.0000	0.00	0	0
40	3	0	0.00	0.00	0.00	0.0000	0.00	0	0
41	4	0	0.00	0.00	0.00	0.0000	0.00	0	0

As can be seen, some items of the scale did not meet the assumption of monotonicity. This does not support the ordering of respondents in terms of their total scores. Figure 2 shows the visual analysis of monotonicity assumptions for four items (items 5, 12, 17, and 20) of the scale. Each plot includes two parts. The left part shows Item Step Response Function (ISRF), which represents the probability of endorsing a certain category across the latent trait θ, and the right part shows Item Response Function (IRF) for the overall item, which characterizes the relationship between the latent trait and items or response options. The plots confirmed the numerical values of monotonicity analysis and illustrated that ISRFs and IRFs are decreasing across the rest score groups for some items.

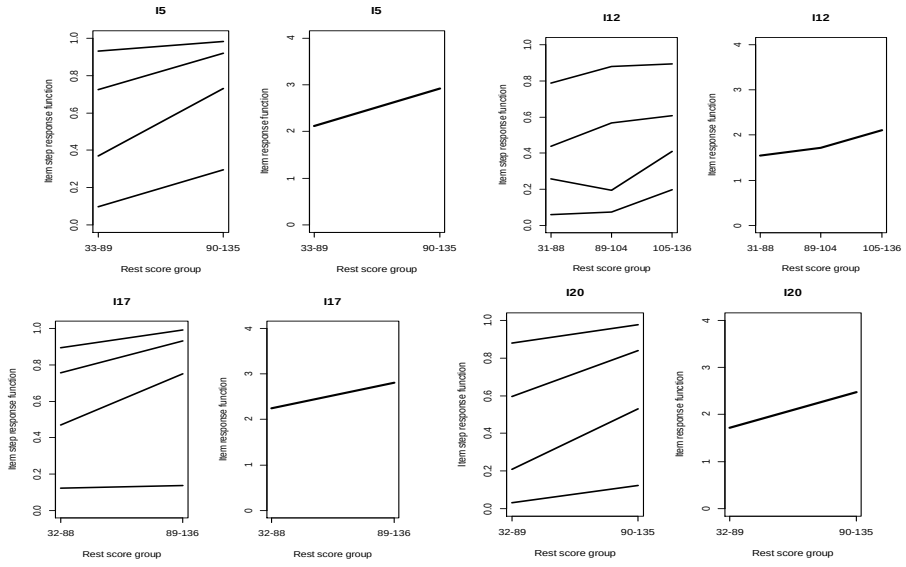

Figure 2. *Monotonicity Plots for Eight Items of the Scale*

Invariant Item Ordering (IIO)

The concept of Invariant Item Ordering (IIO) or non-intersection of IRFs is used to determine if the order of items remains consistent across respondents with different trait levels. IIO adds more meaning to person comparisons based on total scores. For example, a respondent with a higher total score is more likely to have succeeded in an item or endorsed higher categories compared to another respondent. Similarly, two individuals with the same total score are more likely to have answered the same items correctly or endorsed the same categories. To evaluate IIO, Sijtsma and van der Ark (2017) suggested using the coefficient HT. This means that HT < 0.3 indicates insufficient or inaccurate IIO, 0.3 ≤ HT < 0.4 indicates weak IIO, 0.4 ≤ HT < 0.5 indicates medium IIO, and HT ≥ 0.5 indicates strong IIO. Additionally, comparing the IRFs graphically can help determine if two items significantly intersect each other. Table 5 provides the analysis conducted on the scalability of the scale. The second column, labeled "ItemH," displays the scalability coefficient for each item. The third column, "#ac'," indicates the total number of active pairs. Columns four and five, labeled "#vi" and "#vi/#ac," respectively show the total number of violations and the average number of violations per active pair. The sixth column, "maxvi," displays the maximum violation observed. Columns seven and eight, labeled "sum" and "sum/#ac," respectively present the sum of all violations and the average violation per active pair. Columns nine and ten, labeled "tmax" and "#tsig," respectively show the maximum test statistic and the number of significant violations. The last column, labeled "crit," is a weighted sum of other elements such as 'itemH' and '#ac'. The 'crit' value can be used to assess the effect size of IIO violation, with a high value indicating poor items. According to Molenaar and Sijtsma (2000) and van Schuur (2011), a 'crit' value of 0 is perfect, Crit < 40 suggests a minor violation, 40 ≤ Crit < 80 indicates a nonserious violation that requires review, and Crit ≥ 80 signifies a significant or serious violation.

As can be illustrated in Table 5, the H_T value was 0.135, suggesting that item ordering is inaccurate. There were 27 items which violated IIO (column #vi), but only six items (7, 11, 13, 37, 39, and 41) had significant violations (column #tsig). For example, item 41 had 10 violations, that is, the IRF for this item intersected with the IRF of ten other items, but only two of these violations were significant. Similarly, the IRF for item 36 intersected with the IRF of nine other items, but none of them was significant.

Table 5

The Summary of IIO Analysis for the Scale

Items	ItemH	#ac	#vi	#vi/#ac	maxvi	sum	sum/#ac	tmax	#tsig	crit
38	0.26	47	0	0.00	0.00	0.00	0.0000	0.00	0	0
39	0.27	42	2	0.05	0.32	0.46	0.0109	1.68	1	78
7	0.21	49	1	0.02	0.32	0.32	0.0065	1.68	1	72
19	0.25	47	1	0.02	0.14	0.14	0.0030	0.82	0	30
3	0.26	44	0	0.00	0.00	0.00	0.0000	0.00	0	0
4	0.26	45	0	0.00	0.00	0.00	0.0000	0.00	0	0
40	0.26	49	0	0.00	0.00	0.00	0.0000	0.00	0	0
5	0.21	47	0	0.00	0.00	0.00	0.0000	0.00	0	0
17	0.23	46	0	0.00	0.00	0.00	0.0000	0.00	0	0
9	0.24	42	0	0.00	0.00	0.00	0.0000	0.00	0	0
16	0.25	43	0	0.00	0.00	0.00	0.0000	0.00	0	0
2	0.29	48	1	0.02	0.18	0.18	0.0037	0.96	0	33
6	0.27	46	0	0.00	0.00	0.00	0.0000	0.00	0	0
18	0.25	48	1	0.02	0.16	0.16	0.0034	0.89	0	33
30	0.26	47	1	0.02	0.14	0.14	0.0031	0.76	0	30
11	0.33	48	8	0.17	0.36	1.61	0.0336	2.05	1	122
1	0.25	49	4	0.08	0.24	0.72	0.0147	1.31	0	66
29	0.24	49	2	0.04	0.21	0.35	0.0070	1.15	0	48
22	0.26	48	2	0.04	0.22	0.38	0.0079	1.20	0	49
13	0.28	44	8	0.18	0.36	1.85	0.0420	2.00	2	142
41	0.14	47	10	0.21	0.36	2.19	0.0465	2.05	2	158
37	0.21	51	4	0.08	0.31	0.90	0.0176	1.82	1	93
15	0.26	48	3	0.06	0.19	0.54	0.0112	1.23	0	53
23	0.23	50	3	0.06	0.22	0.55	0.0110	1.26	0	58
14	0.25	42	1	0.02	0.18	0.18	0.0043	1.14	0	37
26	0.23	56	2	0.04	0.15	0.29	0.0051	0.83	0	38
21	0.25	42	4	0.10	0.23	0.69	0.0164	1.36	0	68
8	0.21	51	2	0.04	0.15	0.30	0.0059	0.81	0	40
10	0.25	50	3	0.06	0.15	0.41	0.0081	0.88	0	44
20	0.23	43	3	0.07	0.25	0.63	0.0145	1.40	0	66
36	0.14	49	9	0.18	0.25	1.73	0.0353	1.40	0	110
32	0.16	46	5	0.11	0.22	0.88	0.0192	1.27	0	77

27	0.20	41	1	0.02	0.13	0.13	0.0033	0.71	0	32
25	0.21	46	1	0.02	0.18	0.18	0.0040	1.30	0	40
24	0.25	49	1	0.02	0.16	0.16	0.0034	0.91	0	33
34	0.12	47	3	0.06	0.18	0.48	0.0102	1.05	0	57
33	0.14	44	0	0.00	0.00	0.00	0.0000	0.00	0	0
35	0.18	53	0	0.00	0.00	0.00	0.0000	0.00	0	0
12	0.12	44	0	0.00	0.00	0.00	0.0000	0.00	0	0
28	0.14	50	0	0.00	0.00	0.00	0.0000	0.00	0	0
31	0.15	47	0	0.00	0.00	0.00	0.0000	0.00	0	0

*H_T = 0.135

Figure 3 shows the IRFs for two item pairs, for example, items 5/31 and 39/19. As illustrated, the IRFs of item-pairs 5/31 do not intersect; however, for item-pairs 39/19 the IRFs intersect.

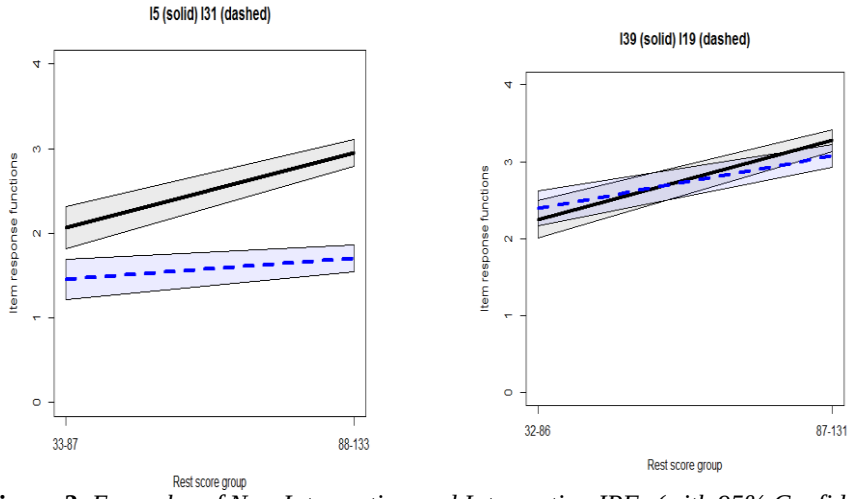

Figure 3. *Examples of Non-Intersecting and Intersecting IRFs (with 95% Confidence Intervals) for Six Item Pairs*

The researchers also employed the backward selection method to eliminate items that violated the assumption of Item-Item Overlap (IIO). In cases where two or more items had an equal number of violations, the item with the lowest scalability, as outlined by Ligtvoet et al. (2010), was removed from the analysis. As demonstrated in Step 1 of Table 6, the results of backward selection showed that six items (7, 11, 13, 37, 39, and 41) have violated IIO and should be removed. To delete items, one item at a time was removed in

iterative steps because "IIO violations of other items may be influenced by the inclusion or exclusion of any particular item" (Stochl et al., 2012, p. 8).

Table 6
The Results of the Backward Item Selection Method

Items	Step1	Step2	Step3	Step4
38	0	0	0	0
39	1	1	1	0
7	1	1	1	NA
19	0	0	0	0
3	0	0	0	0
4	0	0	0	0
40	0	0	0	0
5	0	0	0	0
17	0	0	0	0
9	0	0	0	0
16	0	0	0	0
2	0	0	0	0
6	0	0	0	0
18	0	0	0	0
30	0	0	0	0
11	1	0	0	0
1	0	0	0	0
29	0	0	0	0
22	0	0	0	0
13	2	1	0	0
41	2	NA	NA	NA
37	11	1	NA	NA
15	0	0	0	0
23	0	0	0	0
14	0	0	0	0
26	0	0	0	0
21	0	0	0	0
8	0	0	0	0
10	0	0	0	0

20	0	0	0	0
36	0	0	0	0
32	0	0	0	0
27	0	0	0	0
25	0	0	0	0
24	0	0	0	0
34	0	0	0	0
33	0	0	0	0
35	0	0	0	0
12	0	0	0	0
28	0	0	0	0
31	0	0	0	0

After removing the items violating the IIO, the quality of the 35-item scale was analyzed. In the second round of the analysis, items 20 and 32 violated the IIO and were then removed. The total scalability of the 33-item scale was 0.222. Most of the item pairs were positive, and item scalability coefficients were mostly above the cut-off value of 0.30. The results of IIO and the backward selection method also showed that although there were twelve items which have violated IIO, they were not significant. This indicates that after removing the eight items, the Monotone Homogeneity Model (MHM) and the Double Monotonicity Model (DMM) fitted well to the data.

To investigate the second research question concerning the reliability of the newly created inventory for assessing the intercultural competence of EFL learners, four distinct reliability coefficients were evaluated. These included Mokken scale (MS) reliability, denoted as ρ (Mokken, 1971), Cronbach's alpha (Cronbach, 1951), Lambda-2 (Guttman, 1945), and the latent class reliability coefficient (LCRC; van der Ark et al., 2011). As demonstrated in Table 7, all the values were greater than 0.89, suggesting the high reliability of the scale.

Table 7
Reliability Indices for the 33-Item Scale

Reliability Index	**MS**	**alpha**	**Lambad-2**	**LCRC**
Value	0.895	0.892	0.895	0.904

Discussion

The current study aimed at developing and validating an inventory on EFL learners' intercultural competence so as to investigate to what extent EFL learners in Iran are aware of cultural differences. The results showed that the new instrument possessed high reliability and validity estimates. The results of this study are consistent with prior research conducted by Duisembekova (2021) conducted research on the creation and validation of a new instrument designed to investigate the beliefs of English language student teachers regarding intercultural communicative competence in Turkey. They detailed the process of instrument development and validation, along with a concise review of intercultural competence literature. Their study resulted in a reliable inventory of 34 items with a perfect reliability value of 0.925. Furthermore, Tabatabaee-Yazdi and Baghaei (2022) developed a questionnaire with four dimensions (cognitive, metacognitive, motivational, and behavioral intelligence) to examine the validity of the Persian Intercultural Intelligence Scale using the Rasch rating scale model. Their findings indicated that all items fit the Rasch model without any misfitting items, and the response scale categories were well-differentiated and did not require modification.

The current study aimed to create and validate an inventory assessing the intercultural competence of Iranian EFL learners, utilizing Mokken scale analysis. The study demonstrated that the new instrument exhibited strong reliability and validity. The analyses focused on establishing the instrument's structure and confirming the effectiveness of a five-point response scale. The findings of the study, along with the examination of item characteristics and response scale quality, provided confirmation of the scale's internal validity. According to the results, after the validation process, out of a total of 41 items, 33 items remained. The behavioral and skills construct (items 1-16), the cognitive and metacognitive construct (items 17-31), the motivational construct (items 31-33). The study results and examination of item characteristics and response scale quality confirmed the internal validity of the scale.

This study has made significant strides in recognizing the importance of intercultural competence in education by introducing a reliable and validated version of the intercultural competence scale through self-developed items. Consequently, this instrument can serve as a valuable research tool for investigating the intercultural competence of EFL learners in educational research contexts. Furthermore, the potential adaptation and utilization of this instrument can stimulate research focused on analyzing learners' beliefs

concerning intercultural competence. While there remains a substantial need for further research, both conceptually and statistically, this intercultural competence scale can serve as a foundational resource for comparing findings across various studies and research environments. By utilizing the present intercultural competence scale, researchers can gain insights into the value added by intercultural competence, ultimately contributing to a better understanding of this aspect of education. This information can benefit teachers, curriculum developers, and education researchers, as intercultural competence plays a crucial role in the affective domain of education.

The developed and validated inventory, specifically tailored to the Iranian context, can find practical applications in schools and educational institutions. Consultants and school counselors can employ this questionnaire to evaluate the intercultural competence of EFL learners. The resulting evaluations can guide educational policymakers in making informed decisions to enhance learners' intercultural competence within diverse educational systems. Given these perspectives, the presence of this inventory in the Iranian context, with its strong internal consistency, internal validity, and construct validity, holds significant promise for advancing research in language learners' intercultural competence.

Building upon these findings, the researchers offer several recommendations for future research that can greatly impact educational settings. Subsequent studies could assess the applicability of the inventory in other cultural and contextual settings, with modifications tailored to the specific needs and characteristics of the target population and context. Additionally, broadening the participant pool to include language learners in diverse cultural and linguistic settings, including ESL contexts, can enrich the reliability and validity of the inventory while also fostering more intercultural-educational research. Given the complexity of intercultural competence, further exploration of intervening variables related to intercultural competence is also warranted.

Declaration of interest: none

References

Alijanian, F., Mobini, F., & Ghasemi, P. (2019). The correlation between Iranian EFL learners' intercultural sensitivity, vocabulary knowledge, and English language proficiency. *Issues in Language Teaching, 8*(2), 109-135.

Altshuler, L., Sussman, N. M., & Kachur, E. (2003). Assessing changes in intercultural sensitivity among physician trainees using the intercultural

development inventory. *International Journal of Intercultural Relations, 27*(4), 387-401.

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

Briones, C. N., & Ramos, A. M. (2021). Revitalizing Conversations: Lessons from and About the Production of Intersubjective and Intercultural Knowledge. *Repositorio Institucional CONICET Digital*. DOI:10.33524/cjarv21i3.505

Brooks, N. (1968). Teaching culture in the foreign language classroom. *Foreign Language Annals, 1*(3), 204-217.

Brown, P. (2020). Language as a model for culture: Lessons from the cognitive sciences. In *Anthropology Beyond Culture* (pp. 169-192). Routledge.

Chao, T.-c. (2013). A diary study of university EFL learners' intercultural learning through foreign films. *Language, Culture and Curriculum, 26*(3), 247-265.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika, 16*(3), 297-334. DOI: 10.1007/BF02310555

Cross, T. L. (1989). *Towards a culturally competent system of care: A monograph on effective services for minority children who are severely emotionally disturbed.* CASSP Technical Assistance Center Georgetown, University Child Development Center. Washington, DC.

Duisembekova, Z. (2021). Beliefs about Intercultural Communicative Competence: The Development and Validation of a New Instrument. *International Journal of Instruction, 14*(2), 103-116.

Geertz, C. (2000). The world in pieces: Culture and politics at the end of the century. *Available Light: Anthropological Reflections on Philosophical Topics, 2*, 18-263.

Gray, J. S., Connolly, J. P., & Brown, M. A. (2019). Measuring intercultural knowledge and competence in college essays: Does a performance-based rubric have construct validity? *Studies in Educational Evaluation, 62*, 142-148. DOI: 10.1016/j.stueduc.2019.05.007.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*(4), 255-282. DOI: 10.1007/BF02288892.

Hofstede, G. (2003). What is culture? A reply to Baskerville. *Accounting, Organizations and Society, 28*(7-8), 811-813.

Kim, M. K. (2019). Project-based learning experience in the construction of intercultural knowledge. *Modern English Education Society, 20*(2), 1-18. DOI:10.18095/mee.2019.20.2.1

Kramsch, C. (2014). Language and culture. *AILA Review, 27*(1), 30-55.

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*(4), 578-595. DOI: 1177/0013164409355697

Merrouche, S., & Abderrahim, F. (2006). The place of culture in the teaching of English in the Algerian middle and secondary school.

Mokken, R. J. (2011). *A theory and procedure of scale analysis: With applications in political research* (Vol. 1): Walter de Gruyter. DOI: 10.1515/9783110813203.

Molenaar, W., & Sijtsma, K. (2000). *MSP5 for windows user's manual*. iec ProGAMMA.

Muir, P. (2007). Toward culture: Some basic elements of culture-based instruction in China's high schools. *Sino-US English Teaching, 4*(4), 38-43.

Rangriz, S., & Harati, M. (2017). The relationship between language and culture. *Journal of Applied Linguistics and Language Research, 4*(6), 209-213.

Schat, E., van der Knaap, E., & de Graaff, R. (2021). The development and validation of an intercultural competence evaluation instrument for upper secondary foreign language teaching. *Intercultural Communication Education, 4*(2), 137-154. DOI: 10.29140/ice.v4n2.432.

Schwieter, J. W., Jackson, J., & Ferreira, A. (2021). When 'domestic' and 'international' students study abroad: reflections on language learning, contact, and culture. *International Journal of Bilingual Education and Bilingualism, 24*(1), 124-137. DOI:10.1080/13670050.2018.1447545.

Seelye, H. N. (1993). *Teaching culture: Strategies for intercultural communication* (Vol. 10): National Textbook Company.

ijtsma, K., & van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology, 70*(1), 137-158.

Sijtsma, K., & van der Ark, L. A. (2017). A tutorial on how to do a Mokken scale analysis on your test and questionnaire data. *British Journal of Mathematical and Statistical Psychology, 70*(1), 137-158.

Singhal, M. (1997). The Internet and foreign language education: Benefits and challenges. *The Internet TESL Journal, 3*(6), 107.

Stochl, J., Jones, P. B., & Croudace, T. J. (2012). Mokken scale analysis of mental health and well-being questionnaire item responses: a non-parametric IRT method in empirical research for applied health researchers. *BMC Medical Research Methodology, 12*(1), 1-16. DOI: 10.1186/1471-2288-12-74.

Straat, J. H., Van der Ark, L. A., & Sijtsma, K. (2014). Minimum sample size requirements for Mokken scale analysis. *Educational and Psychological Measurement, 74*(5), 809-822.

Svicher, A., Di Fabio, A., & Gori, A. (2022). Decent work in Italy: A network analysis. *Australian Journal of Career Development, 31*(1), 42-56.

Tabatabaee-Yazdi, M., & Baghaei, P. (2022). Persian Translation and Rasch Model-Based Validation of an Intercultural Intelligence Scale. *International Journal of Society, Culture & Language, 10*(1), 71-82. DOI: 10.22034/ijscl.2021.245308.

Team, R. C. (2020). RA language and environment for statistical computing, R Foundation for Statistical. *Computing*.

Tran, T. Q., & Duong, T. M. (2018). The effectiveness of the intercultural language communicative teaching model for EFL learners. *Asian-Pacific Journal of Second and Foreign Language Education, 3(6)*, 1-17.

van der Ark, L. A., van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement, 35*(5), 380-392. DOI: 10.1177/0146621610392911.

Van Schuur, W. H. (2011). *Ordinal item response theory: Mokken scale analysis:* Sage. Retrieved from https://cran.r-project.org/web/packages/mokken/

Wright, B., D., & Stone, M. H. (1979). *Best test design. Rasch Measurement*. Chicago, IL: ERIC

Zhang, X., & Zhou, M. (2019). Interventions to promote learners' intercultural competence: A meta-analysis. *International Journal of Intercultural Relations, 71*, 31-47. DOI: 10.1016/j.ijintrel.2019.04.006

Biodata

Mahboubeh Akbari holds a BA in English translation from Payam Noor Qochan University in Mashhad. she also has a MA in Teaching English as a Foreign Language from Tabran Institute of Higher Education, Mashhad, Iran. She has 18 years of experience in teaching English, which includes ten years of teaching in Mashhad schools and eight years of teaching in language schools in Mashhad. In 2005, she received a degree in Microsoft Networks Engineering from the UAE.

Mona Tabatabaee-Yazdi is an assistant professor, in the English Department, at Tabaran Institute of Higher Education, Mashhad, Iran. She holds a PhD in Teaching English as a Foreign Language. She has taught English as a Foreign language at several universities and language schools in Iran for 15 years. Her research interests are in Language Assessment and

Teachers Professional Development. Her current focus is on the construction and validation of educational and psychological tests using DCM models. She has participated in post-doctoral research at Friedrich Schiller University of Jena, Germany (2017-2018). She is also the managing editor of the open access *International Journal of Language Testing* which is a scholarly double-blind peer-reviewed international scientific journal published biannually in October and March.