# Validity argument of an Interactional Competence checklist: Perspectives and many-facets Rasch analysis

Zahra Montasseri & Alireza Ahmadi
Shiraz University, Iran

The continued scrutiny of interactional competence (IC) as a sociolinguistically-based construct is gaining momentum in L2 assessment. Previous contributions have emphasized the multi-faceted nature of IC in operationalization and assessment. Among the few assessment tools developed for IC, May et al.'s (2020) checklist is one of the most comprehensive. This study evaluated the decision inference of this IC checklist based on Knoch and Chapelle's (2017) argument-based approach to validation. Several sources of data, including test score data, semi-structured interviews, and expert reviews were used to investigate the proposed assumptions. Four raters evaluated 62 upper-intermediate and advanced test-takers' paired and group oral performances using the adapted IC checklist. Furthermore, five experts in the field of language assessment and 10 language teachers participated in the study. To analyze the data, MFRM analysis through FACETS and thematic analysis were employed. The results provided partial evidence for the assumptions, thus leading to an overall partial support to the decision inference. The IC checklist showed to be a valid instrument for differentiating between test-takers at different IC levels, interpreting scores, and informing teaching and learning. The major drawback of the checklist was associated with a lack of descriptive levels and numerical values.

**Keywords:** interactional competence, argument-based validation, oral performance assessment, low-stake tests, learning-oriented assessment

Email address for correspondence: s.zmontasseri@rose.shirazu.ac.ir

## Introduction

Much of the evaluation and assessment in language classrooms revolve around individual performance rather than the collective efficiency of learners (Walsh, 2012). This focus prioritizes the accurate production of linguistic forms and fluency at the individual level. Speaking tests frequently assess accuracy, fluency, and vocabulary range, rather than the negotiation and co-construction of meaning through interaction (Uludag et al., 2022).

Several factors contribute to this prevailing trend in language classrooms. One reason is the convenience of teaching and testing solo performance, as it is simpler and less time-consuming to design tasks for individuals compared to joint tasks involving multiple participants. Consequently, despite the inclusion of task-based approaches in current materials, learners are not sufficiently trained to engage effectively in interactive discourse. However, effective communication beyond the classroom necessitates collective interaction and the construction of shared meaning. Previously, language proficiency was primarily measured through discrete-point tests, focusing on grammar, vocabulary, and phonology (Lado, 1961). In the 1970s, communicative language approaches introduced a shift toward performance assessment. Pair and group work activities were incorporated into language classrooms with the communicative turn in language teaching, and performance became integrated into communicative-based assessments (Vidaković & Galaczi, 2013). Despite the prevalence of communicative competence models in language testing, criticisms were raised for their cognitive focus and excessive emphasis on individual language users (McNamara, 1996).

Scholars such as McNamara (1996) and Chalhoub-Deville (2003) have proposed a shift in language assessment towards a model that emphasizes the interaction between speakers and their interlocutors. This model highlights concepts like co-construction and interactional competence (IC). As a result, IC has been increasingly integrated into teaching and testing contexts (Kley, 2015). Walsh (2012) predicts that IC will be recognized as the fifth skill, alongside listening, speaking, reading, and writing, in language assessment. He asserts that both teachers and learners should acquire knowledge of how IC is formed and achieved, as interlocutors have varying levels of

competency in joint meaning-making. This understanding promotes dynamic interactions in the classroom and creates an enhanced learning environment.

There has been a growing emphasis on evaluating actual performance and the adaptability of resources in specific situations. As a result, co-constructed discourse has gained prominence in language teaching and learning. Language testers have recognized the importance of incorporating an intercultural perspective into language assessment due to the social aspect of face-to-face interaction. Despite the attention given to IC in language assessment, there is a lack of comprehensive conceptualization and operationalization of this construct (Nakatsuhara et al., 2018). While there have been some rubrics addressing IC, none have fully operationalized all its aspects, nor have assessment tools been solely dedicated to this concept (Lam et al., 2023). Therefore, there is a need for an instrument that can both operationalize IC and provide appropriate feedback to participants.

Recently, May et al. (2020) developed a detailed checklist for assessing IC, aiming to provide learners with feedback focused on their strengths and weaknesses in IC performance. This checklist assists teachers in teaching interactional skills by providing an instrument to deliver feedback on IC skills systematically. It also provides empirically-based IC features that can be integrated into English language teaching materials.

Given the importance of how teachers utilize the IC checklist, its descriptions, and the feedback given to learners, whether in the context of IC teaching or classroom-based assessment, it is essential to understand the usefulness of the checklist and its impact on enhancing learners' IC skills. It is crucial to investigate how different stakeholders interpret and utilize the outcomes of the instrument, taking into account the consequences of the checklist. Therefore, this study aims to investigate the intended use of May et al.'s IC checklist through an argument-based approach to validation.

## Literature review

The complexity of IC and the difficulties in defining and specifying it are evident in various

aspects. A thorough examination of the existing literature reveals multiple definitions of IC, broadly categorized into macro-level, which refers to the meaningful joint construction of interaction considering sociocultural and pragmatic aspects of speech events, and micro-level, which involves the management of topics and turns, pauses, repairs, and non-verbal or body language (Galaczi & Taylor, 2018).

One of the challenging questions about IC is the comprehensiveness of its definition for teaching, learning, and assessment, or what Galaczi and Taylor (2018) describe as construct fine-tuning. Another issue concerns how well the conceptualization of the construct incorporates its central elements. It is important to acknowledge that some of the criteria for assessing IC are culture-sensitive and may not be easily evaluated. Galaczi and Taylor (2018) suggest that all of these features contribute to developing IC skills and that further exploration into operationalizing them in assessment scales would be beneficial. An additional challenge highlighted by Galaczi and Taylor (2018) is the scalability of the criteria, particularly for aspects that are difficult to measure, such as genre awareness. They argue that the difficulties in scalability should not discourage attempts to assess these constructs.

The communicative nature of IC, which requires the involvement of at least two participants for successful completion and evaluation, has presented a significant challenge for evaluators in terms of assigning scores (Liubashenko & Kavytska, 2020). The challenge arises from the difficulty in attributing an individual score to a participant based on their contribution to a collaborative activity involving multiple language users and the subsequent interpretation of this score (Swain, 2001; McNamara, 1997).

Chalhoub-Deville and Deville (2005) assert that in the case of a collectively constructed interaction, it becomes unnecessary to assign individual scores to evaluate each participant separately. Several scholars have suggested the development of evaluation scales that incorporate both shared and individual scoring systems (May, 2009; Ramazani et al., 2019; Taylor & Wigglesworth, 2009). However, some researchers have contested the idea of a shared scoring scale (e.g., Nakatsuhara, 2013). Nevertheless, this remains an issue that requires further empirical investigation.

In recent years, efforts have been made to operationalize IC through various rating scales including the Cambridge English General English Tests, the Test of English for Academic Purposes (TEAP), the Kanda English Proficiency Test (KEPT), and Trinity's Integrated Skills of English (ISE) Speaking and Listening Test. Apart from the limited number of scales developed so far, May et al. (2020) designed a checklist which appears to be the most comprehensive one dedicated solely to assessing IC. Checklists are list-like evaluation tools entailing a list of specific criteria and a place to mark the absence or presence of the characteristics. According to Brookhart (2013, p. 77), checklists "break an assignment down into discrete bits … [which] clarifies what is required for the assignment". Unlike rating scales, which are appropriate for assigning scores, checklists are appropriate for low-stakes decisions like classroom performances as they focus on learners' points of strength and weaknesses (Stevens & Levi, 2005). May et al. (2020) developed a checklist for assessing IC that encompasses the key aspects of learning-oriented assessment. The authors identified several macro themes, such as initiating discussions and introducing new ideas, responding to one's partner, maintaining and developing the interaction, negotiating towards a desired outcome, providing necessary support, interactive listening, body language, and rater reflection, each including several micro themes.

In summary, the existing literature contains numerous studies regarding the concept of IC, with many researchers striving to comprehend and assess this multifaceted construct from both theoretical and operational perspectives; however, challenges persist in defining and evaluating IC. One comprehensive tool for IC assessment is the checklist developed by May et al. (2020), although its empirical validation remains untested. Therefore, the present study seeks to validate May et al.'s (2020) IC checklist utilizing an argument-based validation framework, with a specific focus on the *decision* inference.

Prior to proceeding further, it is necessary to first elucidate the conceptual distinctions between decision inference outlined by Knoch and Chappelle (2017) and the nature of decision-making within the context of the present study. Decision inference as described by Knoch and Chapelle (2017, p. 17) claims that "decisions made based on the estimates

of the quality of the performance are appropriate and well communicated." To warrant the inference, they further state that "Checklist outcomes are suitable for test users and allow for appropriate decision-making."

In the present study, however, we did not intend to address the universal applicability of the checklist. Rather, we assessed the possibility of decisions within the language center (see the Method section) in general and in-class performances in particular, including learners' longitudinal IC development, reports to their parents, their learning portfolio, and the possibility of passing on to the next level. Therefore, the nature of the decisions in this study differs somewhat from the decision inference described by Knoch and Chapelle (2017), as our focus was on using the checklist results to guide learning-oriented decisions within the language program context, rather than for high-stakes assessment purposes.

It is worth mentioning that this checklist is made for providing learners with learning-oriented feedback on their IC skills within classroom scope. Therefore, assessing whether the checklist is appropriate for decision-making purposes, especially in high-stakes tests, goes beyond the primary aim of it. To address these objectives, the following research questions are formulated within the context of Knoch and Chapelle's (2017) argument-based framework of validation:

1. Can the IC checklist effectively differentiate test-takers into strong and weak levels of IC suitable for decision-making?
2. Does the layout of the IC checklist enhance appropriate score reporting and accountability and enable users to make informed decisions?
3. Are IC checklist users able to comprehend the rating scale and associated feedback to make appropriate decisions?

It is prudent to provide definitional clarity on several key concepts that may carry distinct meanings within the specific context of this study, as compared to their more general usage. In general, terms such as ‘decision-making’, ‘performance reporting’, and ‘accountability’ often evoke associations with high-stakes assessment context. However,

in the current investigation, the application of these concepts is oriented towards more formative, learning-centered applications of the IC checklist.

Regarding RQ1, 'level' does not refer to scoring levels typically specified in assessment rubrics. Instead, the RQ explores how students' interactional skills can be illuminated during preparation courses for interactions. In this context, 'levels' refers to assessing whether students are weak or competent enough to handle interactional tasks; therefore, 'levels' is centered around determining learners' weaknesses or strengths rather than quantifying them with a detailed scale.

Moreover, the intended use of the IC checklist is to provide diagnostic feedback to learners, focusing on identifying their strengths and weaknesses, rather than directly providing scores to them. However, considering the two additional purposes of the checklist, which are to "enhance examiners' confidence in awarding IC scores [and] to provide additional scoring validity evidence" (May et al., 2020, p. 52), it is important to note that RQ2 aims to explore whether the checklist can enhance score accountability (i.e., raters assigning scores, teachers interpreting scores, and learners understanding scores) in similar contexts where learners' IC is assessed.

## Method

### Research design and theoretical framework

The present study employed a convergent parallel design, consisting of two distinct phases. In this design, the researchers adopted concurrent timing to gather qualitative and quantitative data simultaneously, while ensuring that each data strand remained independent in terms of collection and analysis. The data from interviews with experts, teachers, learners, and raters were collected separately from the quantitative phase of the study, without any direct link between the two. Conversely, findings obtained from the many-facets Rasch analysis were presented to assess the instrument's ability to discriminate among students at different IC levels. The *decision* inference process involves justifying the intended decisions based on test-takers' performance and test

scores using IC checklist are appropriate and well-communicated to the stakeholders. The warrant for this inference is to support the claim that the IC checklist outcomes are suitable for test users and allow for informed decision-making. Accordingly, three assumptions regarding (1) differentiation of test-takers into levels, (2) appropriate score reporting, and (3) interpretation of scores were made for which several sources of backing, i.e. expert review, MFRM, and interviews with test users were carried to (see Table 1 for a review of the research design).

**Table 1.** Overview of research design

| Assumptions | RQ's | Sources for backing | Participants | Procedure/Analysis |
|---|---|---|---|---|
| 1. IC checklist differentiate test-takers into weak and strong levels needed for decision-making | Differentiation of test-takers into IC levels | ✓ Expert review of IC checklist<br>✓ Statistical analysis (MFRM) | ✓ Sixty-two EFL learners<br>✓ Four raters<br>✓ Five experts | ✓ EFL learners participated in paired and/or group oral tasks.<br>✓ Raters scored performances using IC checklist<br>✓ Experts reviewed the checklist<br>✓ Scores were analyzed in FACETS |
| 2. IC checklist enhances score reporting and accountability and enables users to make informed decisions | Appropriate-ness of IC checklist for score reporting | ✓ Expert review of the IC checklist<br>✓ Interview with test-users | ✓ Five experts<br>✓ Ten EFL teachers<br>✓ Four raters<br>✓ Twelve learners | ✓ Experts reviewed the checklist.<br>✓ Test users were interviewed about the suitability of the checklist and score reporting.<br>✓ Interviews were analyzed through TA with MAXQDA. |
| 3. Test-users can interpret the checklist, scores, and feedback to make appropriate decisions | Interpretation of IC checklist | ✓ Expert review<br>✓ Interview with test-users | ✓ Five experts<br>✓ Ten EFL teachers<br>✓ Four raters<br>✓ Twelve learners | ✓ Experts reviewed the checklist.<br>✓ Test users were interviewed about the suitability of the checklist for interpretation.<br>✓ Interviews were analyzed through TA with MAXQDA. |

## Participants

The participants comprised four different groups of students, raters, teachers, and experts, as described below.

### Students

For the first inference of the study, a total of 62 students (29 males, 33 females) were selected from Shiraz University Language Center in Iran, studying English as a foreign

language (EFL). Their ages ranged from 18 to 32 years, and they were enrolled in advanced, upper-intermediate, and intermediate language courses. In line with the requirement of having the ability to "co-construct interaction in a purposeful and meaningful way, taking into account sociocultural and pragmatic dimensions of the speech situation and event" (Galaczi & Taylor, 2018, p. 5), elementary English language learners were excluded from the sample.

### Raters

Four experienced raters who had familiarity with the IC checklist (see rater training (norming) below for details) and more than seven years of experience in rating oral performance were selected to participate in the study. They were International Development Program (IDP) certified and had experience rating Test of English as a Foreign Language (TOEFL) or International English Language Testing System (IELTS).

### Teachers

As a main group of stakeholders, ten instructors, with at least 10 years of experience in English language teaching, were interviewed for their perception of the IC checklist and its probable decisions. These were the instructors whose students participated in the study and were chosen to be interviewed.

### Experts

To get a professional opinion on the checklist, five experts, three females and two males, who had familiarity with IC and experience in scale evaluation and validation procedures, participated in the study. The experts were university professors of applied linguistics with backgrounds in conducting and publishing research in second language assessment, pragmatics, conversation analysis, and IC.

### Instruments

The instruments used in this study included an IC checklist, protocols for semi-structured interviews, and oral interactional tests. Each is explained in detail below.

## IC checklist

The IC checklist developed by May et al. (2020) was the focus of this study. The checklist is designed in two full and concise versions for providing test users with suitable feedback to make binary judgments. To fulfill the objectives of this study, the latter was utilized, which consists of four main criteria, namely *initiating new ideas*, *keeping the discussion*, *negotiating towards an outcome*, and *using body language appropriately*. Each criterion is specified through several interaction strategies with two levels of feedback, *well done* and *needs more work,* as well as a column to add comments. The concise version of the checklist is an illustrative example of the full version and can be adapted to suit individual learning/teaching situations. Also, the original full version was designed in accordance with Cambridge's B2 First task performances; hence, it is not intended for universal applicability.

## Interview protocols

A series of semi-structured interviews were carried out in the English language, engaging experts, raters, teachers, and learners as participants. These interviews were conducted one on one, via telephone. Prior to each interview, a set of questions was prepared in alignment with a checklist derived from assumptions regarding decision quality, effective teaching and learning, and suggestions for procedural improvement. Throughout the interview process, additional questions were posed accordingly. All interviews were recorded and transcribed to facilitate further analysis. Each interview session lasted approximately 30 to 45 minutes.

## Paired and group oral tests

To evaluate learners' IC utilizing a checklist, two oral examinations were administered focusing on controversial topics (in the context of Iran) and requiring learners to articulate their thoughts in English by expressing their (dis)agreement. During the tasks, learners were provided with a topic and were asked to discuss both views, highlight any advantages or disadvantages of each, and finally express their final decision in form a conversation with other interlocutors. The actual test prompt cards are presented in

Figure 1 below.

| **Task 1** |
| --- |
| Read the following topic and make a conversation with your partner(s). Discuss both views and talk about pros and cons of each. Then, express your opinion. Bring reasons and examples to support your view.<br><br>*Some people argue that when a country is going through a difficult economic, social, or cultural period, individuals should consider immigrating to seek better quality of life elsewhere. On the other hand, there are those who believe that individuals should stay in their home country and contribute to its development.*<br><br>You have 15-20 minutes to discuss the topic. |

| **Task 2** |
| --- |
| Read the following topic and make a conversation with your partner(s). Discuss both views and talk about pros and cons of each. Then, express your opinion. Bring reasons and examples to support your view.<br><br>*Some argue that men and women are equal and should have equal rights, while others believe that fundamental differences between the sexes make them inherently unequal, thus warranting different rights.*<br><br>You have 15-20 minutes to discuss the topic. |

**Figure 1.** Oral tasks prompt cards

As seen above, both tasks had the same structure, instruction, and length but different topics. Both tasks were utilized for group and paired assessment for the sake of connectedness in FACETS analysis. Therefore, both paired and group performances included the two topics. The primary objective of these examinations was to stimulate a lively discussion, wherein participants would employ their linguistic capabilities in conjunction with interactive strategies to sustain the conversation through reasoning and explanation. Whether in paired or group format, each examination lasted between 15 to 20 minutes. In every paired or group performance, individuals of both genders were represented. Of the 62 participants, 22 individuals participated in both group and paired tests with a random order of participation in paired or group tasks. It is noteworthy that the reason behind choosing the tasks in this study was that they are widely used in language classrooms in Iran and the researchers chose them as real-life examples in the local contexts. For the task to be challenging enough, controversial topics (which culturally raise much discussion, argument, and disagreement) in paired and group formats are practiced in speaking courses. Plus, they were carefully selected for this study

based on their broad appeal within the local contexts which do not require specialized technical knowledge.

May et al.'s (2020) was designed based on the speaking component of Cambridge B2 Test paired task (i.e., Part 3 of the test) as a collaborative oral interaction task in which candidates are given a few pictures and three minutes to communicate with the other candidate (two minutes of discussion followed by one minute of reaching a decision on the topic). Interlocutors should be able to express their ideas, explain their opinions, show their (dis)agreement, challenge and evaluate other's ideas, and finally reach a decision through negotiation (Cambridge Assessment English, 2024). Such being the case, the oral tasks in this study were similar to Cambridge B2 paired task in that in both examinations, candidates are asked to discuss ideas, initiate and maintain discussion, and then make a decision. Plus, in both tasks, candidates speak to each other, not to the examiner. Meanwhile, the tasks are different in terms of their length: in Cambridge B2 Test, the task takes 3 minutes while in this study the task takes at least 15 minutes.

Given the circumstances of conducting this research amidst the COVID-19 pandemic, the oral examinations were conducted via Skype. Before the test started, the administrator explained the procedure clearly and showed the prompt on the screen with all cameras on. Once candidates started the conversation, the administrator turned off their camera and microphone with the session being recorded.

## Data collection and analysis procedure

To gather data to substantiate the assumptions of the study, a total of 62 EFL learners participated in paired and group oral tasks. First, forty-two of the participants were randomly assigned to three-member groups, forming a total of 14 groups. Once again, another random 42 learners were paired up, resulting in 21 pairs of interlocutors. Out of the whole 63 participants, 22 individuals took part in both paired and group tasks.

Both paired and group interactions lasted for a duration of 15 to 20 minutes. Subsequently, four trained raters evaluated the learners' performances based on the concise version of the IC checklist as an example of how the full version can be

streamlined. Raters were provided with comprehensive details regarding the checklist and its various components to avoid any potential issues with the instrument (see Rater Training (Norming) below for details).

To enhance the reliability of the ratings, every speaker's performance was evaluated by two raters. To enable the utilization of multi-facets Rasch model (MFRM), the samples were distributed among the raters in a manner that each rater had some common samples with the other three raters. The data obtained from the IC checklist were analyzed using MFRM and a four-facet design, with raters, test-takers, formats of the tasks (i.e., paired vs. group), and items of the IC checklist (e.g. initiating new ideas, develop own ideas, reaching joint decision, body language) as the facets. The numerical values assigned to the checklist categories were as follows: 0 represented “needs more work” and 1 represented “well done.”

In the qualitative phase of the study, five experts reviewed the concise version of the IC checklist and provided their insights through semi-structured interviews. Other stakeholders such as raters, teachers, and learners participated in semi-structured interviews conducted via the telephone. These interviews were recorded and transcribed for analysis.

To analyze the transcribed data, thematic analysis (TA) was employed to organize the data and establish a systematic categorization of themes. The data coding process involved several stages of TA, as explained by Braun and Clarke (2006), including the following:

Phase 1-Familiarizing with the data: at this stage, the researchers immersed themselves into the data by reading and rereading the transcripts actively, analytically, and critically, and then writing reflective log or memos as a preliminary step to coding scheme development.

Phase 2- Generating initial codes: at this stage the systematic analysis of data began in which in vivo codes and labels were assigned to identify certain features related to the research question.

Phase 3-Searching for themes: This stage involved active and dynamic process of generating relatable patterns of response (themes) as levels of abstractions beyond categories at the highest level of data pyramid.

Phase 4-Reviewing potential themes: This stage was concerned with quality checking in which the themes were recursively reviewed across the large data set and primary categories to discover any instances of non-alignment and discard or relocate the codes under another theme to finally come up with a coherent analysis.

Phase 5-Defining and naming themes: At this stage, the themes were allocated clear, unique, and specific names fully indicating the essence of each theme. Themes had a singular focus, and any overlaps were addressed to avoid repetition.

Phase 6-Producing the report: At this final stage, the findings were reported in relation to plausible, convincing, clear, and complex explanations, justifications, and generalizations based on the connections between the themes.

Having said that, Braun and Clarke (2006) suggested a concise 15-item checklist for carrying out an appropriate TA that strengthens the rigor of this approach. The checklist includes five main categories, each with several sub-categories and was used for analyzing the interviews collected from participants in this study. To demonstrate the trustworthiness of TA, researchers double-coded about 30% of the transcript and inter-coder agreement showed a high rate of 91.0%.

Furthermore, MAXQDA (Windows version 2020) was utilized for coding and analyzing the interview data, facilitating data organization and the development of a systematic categorization of themes.

## Rater training (Norming)

To mitigate potential discrepancies among raters and alleviate any biases, a training program was implemented, wherein the criteria for assessment and performance samples were discussed under the guidance of a trainer. Due to the unprecedented circumstances

caused by the COVID-19 pandemic, the training was conducted online.

The training program encompassed three distinct sessions, each lasting approximately two hours, with a one-week interval between each session. In essence, these training sessions comprised the following phases:

1. General overview of IC
2. Familiarization with the IC checklist
3. Familiarization with the learning-oriented feedback
4. Interactive task types
5. Rating practice
6. Discussion and consensus building

During the training sessions, IC along with its historical background and prominent characteristics were thoroughly clarified. A comprehensive explanation of the IC checklist was provided, where each subcategory, item, and level of description were meticulously analyzed and raters were furnished with specific details about the scoring procedure. It is worth mentioning that for a thorough understanding of the checklist's content and its application in evaluating oral performances, both the concise and full versions of the IC checklist were used during these sessions. Following this, the oral tasks and the rationale behind their selection were presented and a minimum of two sample tasks were practiced for each task type, allowing raters to assign scores to the performances based on the provided instructional materials complemented by group discussions and score negotiations. This final stage persisted until raters achieved consensus and felt confident in assigning scores to test-takers.

## Results and Discussion

Our analysis of the interview data identified several themes about the IC checklist (see Table 2).

**Table 2.** Themes extracted from interviews and their frequencies

| # | *Experts* | *N* | *Raters* | *N* | *Teachers* | *N* | *Learners* | *N* |
|---|---|---|---|---|---|---|---|---|
| **1** | Authentic language | 7 | Checklist fairness | 8 | Objectivity | 7 | Changes in studying habits | 14 |
| **2** | High levels of ability | 9 | Learning-oriented feedback | 10 | Positive interactional modifications | 4 | Awareness raising | 11 |
| **3** | Record of progress | 4 | Sub-scales | 9 | Motivation | 10 | Cultural sensitivity | 5 |
| **4** | Detailed dimensions | 12 | Individual vs. group performance | 5 | Real-life situations | 5 | Personality traits | 6 |
| **5** | Lack of descriptive levels | 6 | | | Understandable report | 9 | Vocabulary and grammar | 10 |
| **6** | Curriculum and checklist overlap | 11 | | | | | | |

The table displays four sets of data based on the codes extracted in interviews with groups of participants, each containing several themes with their frequencies. For the sake of brevity, only the major themes are discussed in the following pages in accordance to each research question.

## RQ1. Can the IC checklist differentiate test-takers into IC levels needed for decision-making?

To determine if the IC checklist effectively differentiates test-takers into decision-making levels, expert reviews of the scale and many-facets Rasch analysis were conducted. To this end, the FACETS program was employed to analyze the data and calibrate raters, test-takers, format of the tasks, and the IC checklist items onto the logit scale. A single frame of reference for the interpretation of the results is displayed in Figure 2. The first column on the left is the logit scale (measurement scale).

The second column, oriented positively, displays the estimates of the test-takers' IC with ability measures ranging from +0.9 to -1.3 logits.

The third column represents raters' level of severity in rating with more severe raters positioned lower in the column. According to the figure, raters 2 and 3 are equally positioned as the least severe raters. With a nuance of distance from these two are Raters 1 and 4, respectively. The severity measures range from approximately +0.3 to +0.5 logits, indicating a considerable similarity in the rating of the items.

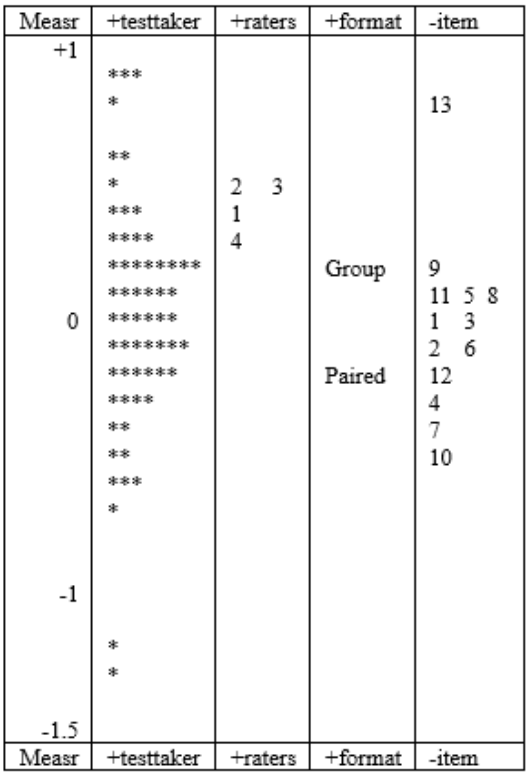


**Figure 2.** Variable map from the many-facets Rasch analysis

Notes. items codes: 1=new ideas; 2=right time for new ideas; 3=language for initiating ideas; 4=develop own idea; 5=develop partner's idea; 6=invite; 7=listen; 8=be collaborative; 9=language for negotiation; 10=joint decision; 11=language for outcome; 12=body language; 13=eye contact

The next column, also positively scaled, indicates the format of the oral task and reveals that test-takers found the paired test (-0.2 logits) more challenging compared to the group test (+0.2 logits).

Lastly, the negatively oriented last column indicates that items of the IC checklist positioned at the top are perceived as more difficult by raters and have higher difficulty measures. The item difficulty measures range from approximately +0.9 to -0.5 logits, with item 13 (eye contact) being notably distinct from the other items.

The measurement reports produced by FACETS offer valuable insights into Pearson-ability measures, difficulty levels of items or task formats within each facet, as well as information on quality control fit statistics (MNSQ and ZSTD). These statistics assess the degree to which each facet fits the model and provide confidence in the associated measures (logits).

As shown in the variable map, there was a wide variation in test-takers' IC with spread logit measures. In addition to the information in the variable map, FACETS presents the measurement results for examinees as seen in the following table.

**Table 3.** Summary of test-takers' measurement report

| Total Score | Total Count | Obsvd Average | Fair (M) Average | +Model Measure | S.E. | Infit MnSq | Infit ZStd | Outfit MnSq | Outfit ZStd | Correlation PtMea | Pt Exp | N | Test-taker |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 21.2 | 35.2 | .60 | .60 | .00 | .37 | 1.00 | -.1 | 1.00 | .0 | .15 | | Mean (Count: 62) | |
| 8.5 | 12.4 | .10 | .10 | .44 | .06 | .07 | .9 | .09 | 1.0 | .53 | | S.D. (Population) | |
| 8.5 | 12.5 | .10 | .10 | .44 | .06 | .07 | .9 | .09 | 1.0 | .53 | | S.D. (Sample) | |

| | | | |
|---|---|---|---|
| Model. Populn: RMSE .38 | Adj (True) S.D. .52 | Separation 1.20 | Strata 1.93 | Reliability .86 |
| Model. Sample: RMSE .38 | Adj (True) S.D. .53 | Separation 1.21 | Strata 1.95 | Reliability .87 |
| Model. Fixed (all same) | Chi-square: 79.2 | d.f.: 61 | Significance (probability): .86 | |
| Model. Random (normal) | Chi-square: 35.6 | d.f.: 60 | Significance (probability): .89 | |

In the above table, the separation index and strata indicate true abilities which show the degree to which test-takers are well-differentiated according to their level of competence. Since the strata value for test-takers is 1.95, it could be asserted that the checklist is able to statistically distinguish between strong and weak performers, or according to the scale, 'well done' and 'needs more work'. Regarding the reliability ($r$=.87), unlike inter-rater reliability which shows that measures are reliably the same, this value indicates the reproducibility of the order of the given measures. According to Green (2013), for the test-taker facet, a higher reliability index (near 1.00) is preferred. Additionally, FACETS provided non-significant values that indicate no significant difference in terms of test-takers' proficiency measure (fixed chi-square=79.2, d.f.=61, $p$=.86) and that the test-takers' IC measures were a random sample from a normal distribution (random chi-square=35.6, d.f.=60, $p$=.89).

Table 4 provides relevant information regarding rater measurement.

**Table 4.** Rater's measurement report

| Total Score | Total Count | Obsvd Ave-rage | Fair (M) Ave-rage | +Model Measure | S.E. | Infit MnSq | Infit ZStd | Outfit MnSq | Outfit ZStd | Estim. Discrm | Correlation PtMea | Pt Exp | N | raters |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 321 | 533 | .60 | .62 | .49 | .09 | 1.02 | -.5 | 1.02 | .6 | .91 | .21 | .24 | 2 | 2 |
| 340 | 559 | .61 | .62 | .49 | .09 | .99 | -.4 | .97 | -.8 | 1.09 | .27 | .24 | 3 | 3 |
| 329 | 533 | .62 | .60 | .42 | .09 | 1.01 | .3 | 1.03 | .7 | .93 | .20 | .23 | 1 | 1 |
| 325 | 559 | .58 | .58 | .34 | .09 | .99 | -.4 | .98 | -.7 | 1.08 | .27 | .25 | 4 | 4 |
| 328.8 | 546.0 | .60 | .61 | .43 | .09 | 1.00 | .0 | 1.00 | -.1 | | .24 | | Mean (Count: 4) | |
| 7.1 | 13.0 | .01 | .01 | .06 | .00 | .01 | .5 | .02 | .8 | | .03 | | S.D. (Population) | |
| 8.2 | 15.0 | .02 | .02 | .07 | .00 | .01 | .5 | .03 | .9 | | .04 | | S.D. (Sample) | |

| | | | |
|---|---|---|---|
| Model. Populn: RMSE .09 | Adj (True) S.D. .00 | Separation .00 | Strata .33 | Reliability(not inter-rater).85 |
| Model. Sample: RMSE .09 | Adj (True) S.D. .00 | Separation .00 | Strata .33 | Reliability(not inter-rater).92 |
| Model. Fixed (all same) | Chi-square: 13.3 | d.f.: 3 | Significance (probability): .59 | |
| Model. Random (normal) | Chi-square: 1.2 | d.f.: 2 | Significance (probability): .55 | |
| Inter-Rater agreement opportunities: 1092 | | Exact agreement: 598.4= 54.8% | Expected: 598.6 – 54.8% | |

The "measure" column indicates that higher measure values associated with a rater correspond to higher levels being assigned to the items. The most lenient raters were Raters 2 and 3, with a severity measure of 0.49 logits, while the most severe rater was Rater 4, with a severity measure of 0.34 logits. Quality-control fit statistics are given that inform about the quality of the data and confidence in the produced measures. These figures contain information regarding information weighted fit statistic (henceforth Infit), outlier-sensitive fit statistic (henceforth Outfit), mean square fit statistics (henceforth MNSQ), and Z-standardized which refers to the t-tests of whether the data fit the model perfectly. For infit and outfit MNSQ values, the acceptable range is between 0.5 and 1.5, while for z-standardized values, it is between +2 and -2 (Linacre, 2021).

As shown in Table 4, the Infit MNSQ for all raters exhibits expected values, indicating consistent use of the rating scale across the items. This confirms the raters' internal consistency, as infit MNSQ is sensitive to unexpected ratings such as carelessness. Both infit and outfit expected values are set at 1.0, implying that the variation in the ratings is neither underfit nor overfit (Myford & Wolfe, 2003). The infit MNSQ values for raters range between 0.99 and 1.2 logits. Notably, infit MNSQ is associated with higher estimation precision and represents a greater threat to measurement compared to outfit MNSQ, which poses less of a threat. Infit is often more significant than outfit in assessing rater fit (Green, 2013). Outfit MNSQ, on the other hand, is more sensitive to unexpected observations on extremely easy or difficult items, random responses, and guessing. The outfit MNSQ figures in the above-mentioned table exhibit a similarity to infit MNSQ

statistics, which is typical for rating scales (Green, 2013). There are no unexpected or borderline outfits among the raters. The z-standardized column also demonstrates that all raters' performance is neither overfitted nor misfit, as their values range between +2 and -2. Additionally, the figures in the "estimated discrimination" column fall within the range of 0.5 to 1.5, indicating reasonable fits to the Rasch model.

Table 4 indicates inter-rater agreement by comparing the observed agreement percentage (exact agree. obs%) in the data to the agreement predicted by the model (exact agree. exp%). Raters exhibit behavior similar to independent experts when the observed percentage is approximately equal to the expected percentage. Conversely, if the observed figures are significantly larger than the expected figures, it suggests that raters may be rating in a similar, predictable manner, possibly compromising their independence. Alternatively, if the observed percentage is lower than the expected percentage, raters' behavior is considered unpredictable (Eckes, 2011). In this study, the exact agreement statistics are either the same or slightly higher than the expected agreement statistics for raters, indicating that they behave like independent experts.

Table 5 suggests the relative difficulty of paired interactions compared to group interactions as the values depicted for the total score, average score, and measured value are smaller for paired interactions. Regarding the quality-control fit statistics, both Infit and Outfit MNSQs are within the acceptable range, i.e., 0.5 to 1.5, and have the ideal 1.0 logits. This in turn confirms the alignment of the group and paired performances with the model. The ZSTD figures stay within the reasonable fit. Also, the chi-square value is 13.3 (df:1, $p$<.001) indicating a statistically significant format effect on test-takers' performances.

**Table 5.** The measurement reports for the Format facet in the MFRM

| Total Score | Total Count | Obsvd Average | Fair (M) Ave-rage | +Model Measure | S.E. | Infit MnSq | Infit ZStd | Outfit MnSq | Outfit ZStd | Estim. Discrm | Correlation PtMea | Pt Exp | N | Format |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 683 | 1092 | .63 | .64 | .16 | .06 | 1.00 | -.1 | 1.00 | -.1 | 1.02 | .26 | .26 | Group | Group |
| 632 | 1092 | .58 | .57 | -.16 | .06 | 1.00 | .2 | 1.00 | -0 | .98 | .20 | .21 | Paired | Paired |
| 657.5 | 1092 | .60 | .61 | .00 | .06 | 1.00 | .1 | 1.00 | -.1 | | .23 | | Mean (Count: 21) | |
| 25.5 | | .02 | .04 | .16 | .00 | .00 | .2 | .00 | .1 | | .03 | | S.D. (Population) | |
| 36.1 | | .03 | .06 | .23 | .00 | .01 | .3 | .00 | .2 | | .04 | | S.D. (Sample) | |
| Model. Populn: RMSE | | | .06 | Adj (True) | | S.D. .15 | | Separation 2.38 | | | Strata 3.50 | | Reliability .85 | |
| Model. Sample: RMSE | | | .06 | Adj (True) | | S.D. .22 | | Separation 3.51 | | | Strata 5.01 | | Reliability .92 | |
| Model. Fixed (all same) | | | Chi-square: 13.3 | | | d.f.: 1 | | | | | Significance (probability): .00 | | | |

Table 6 presents information about the difficulty level of the items and items-model fit. As depicted, item 13 (eye contact) is judged by raters to be the most difficult item and item 10 (joint decision) the easiest one. Considering the fit statistics, all the Infit and Outfit MNSQs stay within the acceptable range. Also, chi-square value is 43.0 (df:1, $p$<.001); i.e., all the items align with what the model expects and they are not equally difficult (i.e., items are significantly different in terms of their difficulty). Another indicator of the suitability of items is no observation of negative values in the estimated discrimination.

**Table 6.** The measurement reports for the Item facet in MFRM

| Total Score | Total Count | Obsvd Average | Fair (M) Average | +Model Measure | S.E. | Infit MnSq | Infit ZStd | Outfit MnSq | Outfit ZStd | Estim Discr | Correlation PtMea | Pt Exp | N | Item |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 131 | 168 | .78 | .79 | .87 | .19 | .95 | -.5 | .92 | -.6 | 1.09 | .30 | .17 | 10 | 10 |
| 108 | 168 | .64 | .65 | .17 | .16 | .99 | -.1 | 1.00 | .0 | 1.03 | .21 | .20 | 7 | 7 |
| 107 | 168 | .64 | .64 | .14 | .16 | 1.00 | .0 | 1.01 | .1 | .99 | .19 | .20 | 4 | 4 |
| 106 | 168 | .63 | .63 | .12 | .16 | 1.01 | .1 | .99 | .0 | .98 | .19 | .20 | 12 | 12 |
| 106 | 168 | .63 | .63 | .12 | .16 | .93 | -1.4 | .92 | -1.4 | 1.36 | .36 | .20 | 3 | 3 |
| 103 | 168 | .61 | .62 | .04 | .16 | .97 | -.6 | .98 | -.4 | 1.15 | .26 | .20 | 6 | 6 |
| 101 | 168 | .60 | .60 | -.02 | .16 | .98 | -.4 | .97 | -.6 | 1.16 | .25 | .20 | 2 | 2 |
| 99 | 168 | .59 | .59 | -.07 | .16 | 1.11 | 2.6 | 1.13 | 2.6 | .17 | -.06 | .20 | 1 | 1 |
| 99 | 168 | .59 | .59 | -.07 | .16 | 1.05 | 1.1 | 1.07 | 1.5 | .60 | .08 | .20 | 11 | 11 |
| 95 | 168 | .57 | .57 | -.17 | .16 | .97 | -.7 | .96 | -.9 | 1.28 | .27 | .20 | 8 | 8 |
| 90 | 168 | .54 | .53 | -.29 | .16 | 1.00 | -.1 | .99 | -.2 | 1.06 | .22 | .20 | 5 | 5 |
| 87 | 168 | .52 | .52 | -.37 | .16 | 1.00 | -.1 | .99 | -.2 | 1.06 | .21 | .20 | 9 | 9 |
| 83 | 168 | .49 | .49 | -.47 | .16 | 1.04 | 1.2 | 1.05 | 1.3 | .50 | .11 | .21 | 13 | 13 |
| 101.2 | 168.0 | .60 | .60 | .00 | .16 | 1.00 | .1 | 1.00 | .1 | | .20 | | Mean (Count: 62) | |
| 11.5 | .0 | .07 | .7 | .32 | .01 | .04 | 1.0 | .06 | 1.1 | | .10 | | S.D. (Population) | |
| 12.0 | .0 | .07 | .7 | .33 | .01 | .05 | 1.1 | .06 | 1.1 | | .11 | | S.D. (Sample) | |

| | | | | |
|---|---|---|---|---|
| Model. Populn: RMSE .16 | Adj (True) S.D. .27 | Separation 1.67 | Strata 2.56 | Reliability .74 |
| Model. Sample: RMSE .16 | Adj (True) S.D. .29 | Separation 1.76 | Strata 2.68 | Reliability .76 |
| Model. Fixed (all same) | Chi-square: 43.0 | d.f.: 12 | Significance (probability): .00 | |
| Model. Random (normal) | Chi-square: 9.4 | d.f.: 11 | Significance (probability): .59 | |

During interviews, the adequacy of the items was confirmed, as the scale was meticulously created to encompass the intricacies of a successful interaction. As expressed by the experts:

> It very much depends on the decision that is to be made about the test-takers. However, generally speaking, yes, to a decent extent because of the detailedness of the dimensions. (Expert C)

> The items on the scale seem to be well-established and, in my opinion, it entails all key features of the IC. (Expert A)

This is supported by the importance of utilizing a framework or model when developing language assessment tools. Without a robust underlying model, item selection would lack principled, consistent, and objective criteria (Lieberman & Michael, 1986). Hence, test designers enhance the likelihood of constructing an objective, systematic, and comprehensive assessment instrument by anchoring the items to a sound theoretical

model.

One theme proposed by test users was that the checklist's insufficient descriptive levels would pose obstacles to future decision-making. They believed that this would hinder the clear classification of test-takers' abilities and make it challenging to make appropriate decisions:

> However, commenting has been limited to 'well done' and 'needs more work'; there is nothing in between and 'needs more work' does not say how much more work. The demand for higher precision becomes more important if a great number of test-takers are involved. (Expert D)

> I believe the number of levels is not enough. We need to specify the levels of the scale, by for example adding at least one more level: 'needs more work', 'competent', 'excellent'. Maybe we can expand the levels later during the refinement. (Expert E)

> The optimum number of levels for a scale is three, so the current scale does not fully differentiate learners on how well or poorly they perform the task. I suggest the following descriptive levels: 'Advanced', 'Intermediate', and 'Novice'. (Expert B)

However, there are reasons for using only two levels in the checklist. IC is a complex construct with challenges in definition, instruction, and assessment, and "its theoretical conceptualization and practical operationalization have not been fully developed in terms of informing the teaching and learning of interactional skills in a comprehensive and user-friendly way" (Nakatsuhara et al., 2018, p. 4). Given the complexity and authenticity of IC and the obstacles in its assessment, test designers have partially integrated it into other oral performance and speaking scales instead of treating it as an independent entity. Previously, the operationalization of IC was limited to a list of interactional elements where raters would indicate the presence or absence of each criterion (Vo, 2019). IC assessment, despite all the efforts, is still in its infancy.

Another possible reason for the absence of descriptive levels is that first, the purpose of this checklist to provide points of strength and weaknesses and second, IC requires higher levels of language proficiency to effectively engage in complex aspects of interaction, such as topic shift, turn-taking strategies, joint utterance creation, and listener involvement. Within this complex network of functions, some aspects are particularly challenging to

assess.

Regarding the MFRM analysis, one possible explanation for raters' independence in scoring could be attributed to the training sessions. Following the training, raters demonstrated greater consistency and a reduction in severity differences. It is important to note that the establishment of too close connections among raters is not recommended to avoid rating dependence. It is claimed that training in a supportive environment leads to more effective results, as raters align themselves more closely with the benchmark (Hamilton et al., 2001). The results also show that test-takers performed better in group discussions compared to paired interactions, indicating that two-way conversations are more challenging.

Regarding the suitability of the items, the results showed that all fit statistics were within an acceptable range in terms of extreme calibration. However, item 13 (eye contact) was perceived as the most challenging item by raters, despite its fit to the model. This lack of consecutiveness suggests that there may be areas of the variable that are poorly defined or insufficiently studied. In this case, no item exhibited an extreme calibration warranting elimination; only *eye contact* appeared to require further investigation. One reason could be the mode of delivery of the oral performance; as mentioned earlier, since this study was conducted during the COVID-19 pandemic when all educational institutions went online, and this study was not an exception, assessing eye contact could have been quite challenging, thus FACETS results might have been impacted. As a result, given that eye contact seemed to be the most difficult item to be assessed, future studies need to address this item in a face-to-face delivery mode.

Some test-takers questioned the inclusion of *eye contact* in the checklist, citing factors such as individual differences, cultural variation, and operational complexities. They believed that eye contact is influenced by personal character and behavior, making it unsuitable for a rating scale. Psychological research has shown a connection between gaze behavior and personality traits (such as nurturing personality, neuroticism, extraversion, and openness), using techniques like eye-tracking and video recording to explore this relationship (Rauthmann et al., 2012). Test-users noted that eye contact can be culture-

specific, with people from different geographical and cultural backgrounds holding diverse and even contradictory perceptions of eye contact. The appropriate level and extent of eye contact during interactions remained unclear in the checklist, posing challenges in its measurement.

Lastly, it is important to explore the relationship between nonverbal behavior (hereinafter referred to as NVB) in general, and specifically eye contact, and proficiency levels. Research suggests that advanced learners demonstrate more instances of NVB, including eye contact, facial expressions, laughter, gestures, postures, and nods, indicating higher levels of engagement and enthusiasm (Banerjee & Plough, 2016). Furthermore, direct eye gaze has been linked to increased comfort and confidence among individuals with higher proficiency levels. The significance of NVB in the co-construction of meaning in high-stakes tests has been supported by Jenkins and Parra (2003). Their microanalysis of recorded interviews led them to conclude that NVB and paralinguistic cues played a crucial role in determining students' success in passing the test. Therefore, eye contact is an essential component of NVB not to be easily omitted from the IC checklist. Instead, it may need to be presented more effectively to accommodate cultural contexts, as extensive eye contact can be considered inappropriate in some Middle Eastern cultures, where this study is situated (Gounaili, 2011). As such, users are encouraged to modify the checklist items as needed to align with local norms and preferences. It is worth mentioning that the concise checklist is an example of how the full version can be streamlined, rather than a one-size-fits-all solution. Therefore, checklist users can exclude or adapt any items to create a version that best suits their specific purposes and contexts.

Altogether, in response to the first research question, the IC checklist demonstrates the capability to differentiate test-takers into distinct levels required for decision-making, as evidenced by test-taker separation index.

### RQ2: Does the IC checklist enhance appropriate score reporting and accountability and enable users to make informed decisions?

To answer this question, interviews were conducted with experts and test users, and key

themes were extracted (see Table 1).

Experts criticized score reporting, noting that it may fail to present feedback to both parties involved in the interaction given that a two-way conversation involves the co-construction of meaning during communication. Successful conversation relies on the mutual efforts and reciprocal shaping of resources by all participants, rather than being solely attributed to one person.

> Much of a successful interaction depends on the partners engaged in the flow of conversation. It seems a little bit unfair to me not to consider how peers affect one another's performances. (Expert A)

> Individual feedback is well-established in this scale, but what about the group-level report? This performance was not individual like writing or listening; it was group-based and needs group report as well. (Expert B)

This poses a challenge in differentiating and evaluating an individual's contribution, casting doubt on the accuracy of score checklists in depicting this phenomenon.

To address this concern, several solutions have been proposed. One approach suggests that raters provide descriptive reports alongside test scores to minimize potential bias in score reporting (Nakatsuhara, 2013). These reports would outline key aspects of an individual's IC and offer suggestions for improvement. Another suggestion is for raters to use *contextualized notes* to explain scores and elaborate on co-constructed discourse (Konzett-Firth, 2020). It is proposed that test-takers be provided with dual scores: one reflecting their abilities and another indicating the aspects of IC that require mutual effort and achievement. This allows raters to compare candidates more effectively and aligns better with the co-construction of communication.

Another theme that emerged from the data is the perceived usefulness of scores by some learners and raters, beyond the basic *well done* versus *needs more work* diagnostic categorization. Related excerpts are given below:

> What learners usually look after is a total score instead of checkboxes because a point system tells them about differential weights and increases their perceptions. A simple score assures them that the teacher is impartial. (Rater B)

> It's good to know which factors of my interaction is fulfilled and which need more work. But I want to know my score on this test. (Learner 12)

> I wonder why this checklist doesn't give us a final score. Our teacher always tells us our score on speaking; she'd better do the same for interaction. (Learner 8)

Criticisms were raised regarding the meaningfulness of scores obtained from the checklist, as it failed to evaluate test-takers' vocabulary and grammar proficiency.

> The layout is appropriate for score reporting to some extent; however, it doesn't really take into account different aspects of language, e.g., range of vocabulary and grammar. Sections should be dedicated to what is expected of learners. Aside from this, the layout is appropriate. (Teacher A)

To address the exclusion of vocabulary and grammar from the IC checklist, it is important to analyze the definition of IC: "The term has been used by different scholars with different shades of meaning in several different areas of second language learning, teaching, and testing" (Young, 2011, p. 426). Such being the case, the consensus among scholars in this field insufficiently addresses the linguistic aspect of performances. As a result, the checklist disregards linguistic elements such as vocabulary and grammar (although Pekarek Doehler (2018) proposed the concept of grammar-for-interaction as a flexible and evolving toolbox that enables individuals to perform actions in ways that are mutually understandable).

Another commonly voiced criticism by test-takers is the checklist's inadequacy in considering personal characteristics unrelated to interactive abilities in L2 but related to personality traits. One specific item that is often seen as unrealistic is the assessment of introversion versus extroversion, which can impact individuals' interactive performance. Introverts, for instance, may not excel at initiating conversations or breaking the ice and may prefer to remain quiet and reserved.

> Many of these skills are rooted in the mother tongue or even the person's personality in conversing with others. For example, a person may not be able to speak Persian without pause and for a long time. (Learner 5)

> …and some of these items are completely related to people's personality traits. Therefore,

> scoring only on these items in the virtual classroom may not be possible. (Learner 11)

This claim holds some truth, as certain introverted traits can have negative effects on interaction. It is important to note that not all aspects of introversion have a detrimental impact on interpersonal communication (Whalen, 2015). Moreover, while personal characteristics may potentially impact the execution or outcomes of the construct, it is not within the scope of an IC checklist to include them in the assessment. Also, the problem of personality traits affecting interaction applies to assessing IC in general and is not limited to the checklist under investigation. It is suggested that teachers and practitioners take into consideration learners' personal characteristics when practicing IC in the classroom. Further research and investigation in this area may provide valuable insights into the connection between personal characteristics and performance on IC assessments.

Taken together, mixed opinions were expressed in response to the second research question which is to a large extent supported as the suitability of the checklist for enhancing accountability of scores.

### RQ3: Are the IC checklist users able to interpret the rating scale and the associated feedback to make appropriate decisions?

One favorable attribute of the IC checklist, as noted by experts, was the provision of feedback that was useful, comprehensible, and readily accessible to test-takers. The provided feedback is highly regarded by both test users and experts due to its practicality. It incorporates meaningful feedback that focuses on learners, learning processes, and learning outcomes (Purpura & Turner, 2014). Related excerpts are given below:

> The feedback given to candidates are meaningful which are essential components of communication. These feedbacks cannot be addressed effectively through digital technologies, hence requires teacher involvement. (Rater D)

> The feedback in the 'Well done' section encourages learners to keep using the features and can enhance their performances. The 'Needs more work' helps candidates identify their problems and improve that aspect of interaction. (Expert B)

> Of course, the precise feedback to individual learners can be adapted by the teacher in accordance with their actual performance. (Teacher F)

> Everything seems all right. The layout seems user-friendly and logical. The descriptions provided in the concise checklist are short and to the point. (Expert E)

> The checklist offers useful feedback to those who need more work. The phrases provided to them are practical and the directions raise their awareness to their weaknesses. Even the comments who achieved 'well done' are worthwhile. (Expert C)

Previously, IC assessment was integrated within oral proficiency tests for language abilities in formal educational evaluation (van Compernolle, 2013), rather than being considered an independent construct. The present checklist has established an autonomous identity for IC by exclusively measuring the construct itself, rather than perceiving it as a tangential byproduct of oral proficiency. Consequently, the resultant scores largely reflect individuals' interactional skills and are intelligible to them. Another factor contributing to the checklist's comprehensibility can be attributed to its testing methodology. Previous research on the evaluation of IC predominantly employed a conversation analysis approach. The current checklist represents a response to the limitations in IC assessment, providing a user-friendly tool for evaluating IC skills (Nakatsuhara et al., 2018).

Evidence from semi-structured interviews with experts and test users suggests that the overlap between checklist content and curriculum content can enhance decision-making (Wilson & Urick, 2021). Opportunity to learn test content refers to the timely instruction of specific content before assessment which can be measured by the time devoted to reviewing and practicing the checklist, as well as the depth of understanding attained in the classroom setting. Teacher-student interaction quality and extent also contribute to creating this opportunity to learn.

> There is nothing blurry in the checklist. The items and sub-scales are clear and understandable. More importantly, the feedback attached to it make interpretation even easier because it specifies where exactly the problem needs to be tackled. (Teacher C)

> The syllabus, textbook, teaching approach, and classroom activities were somehow in line with interactional strategies, so students were familiar with the assessment tasks and what was expected from them. (Teacher G)

Experts have identified three key areas about this assertion. Firstly, the alignment between test content and curriculum content possesses a positive impact on achievement levels. Research has shown that students who are exposed to relevant test content during instructional sessions tend to perform better (Suárez & Gesa, 2019). Secondly, the coverage of material in the assessment is influenced by factors including the nature of the curriculum, teacher supervision, and allocated time for related activities - all important predictors of test performance. In other words, test-takers are more likely to excel if they have received instruction on and are familiar with the assessment material, including its format. Hence, students' preparation with the checklist provides a valuable learning opportunity that enhances future decision-making. Thirdly, the degree of coordination among instructional goals, course objectives, methods, and materials represent another crucial consideration. Research suggests that students' achievement is more assured when various elements within the classroom are presented in a cohesive manner, rather than as separate and discrete components (Vázquez et al., 2020). This implies that optimal performance and maximized learning opportunities occur when there is alignment between the teacher, activities, and curriculum. Therefore, establishing a justifiable alignment between curriculum content and the IC checklist promotes appropriate interpretation of students' performance on the checklist for decision-making.

The checklist was found to be easily interpretable, as supported by additional findings. The language utilized for descriptors was appropriately balanced, avoiding both excessive simplicity and unnecessary technicality, thus aiding in understanding.

> Despite the complexity and multi-faceted nature of IC, the language used in the scale is not twisted at all. I should take my hat off to checklist developers (laughs) for removing the complexity from 'the complex. (Rater B)

> The truth is, teachers and learners may not understand high-level concepts; so, it's better to keep the specifics for the technical audience. Tailoring the language of an instrument to test users is the key to interpretation. (Rater D)

> The first spot to get the information is the levels seen in the checklist. Both teachers and test-takers can explain their strengths and weaknesses. (Rater A)

Experts in the field acknowledged this attribute and provided a multitude of justifications.

Firstly, the checklist demonstrated overall comprehensibility, evidenced in three ways: (1) exhibiting coherence and a logical structure; (2) employing terminology that did not induce confusion; and (3 encompassing descriptors that were deemed relevant for assessing the target construct. The checklist consistently presented coherent and logically organized items, devoid of any unexpected or vague entries. All descriptors were supported by ample references, ensuring that raters could easily make sense of them. Unlike many other scales, the checklist abstained from adopting fuzzy or ambiguous language in its items, except for the previously discussed matter of eye contact. The terms "needs more work" and "well done" were distinctly discernible, offering learning-oriented feedback at the end of the checklist. Importantly, no irrelevant items were identified within the checklist, which could potentially impede interpretation or distort the rating process.

> The feedback gives students an explanation of what they are doing correctly or incorrectly. It is educative in nature and helps them realize their inaccuracies and correct them accordingly. (Expert E)

> If effective feedback is given, it guides learners to adjust specific interactional strategies and become autonomous and self-reflective about their own abilities. Also, it also guides teachers to accommodate learners' needs. (Expert D)

The second issue pertains to the brevity of the checklist, widely appreciated by nearly all interviewees. This characteristic contributes to the checklist's usability for interpretation purposes, particularly considering the time constraints involved in making judgments. Furthermore, raters preferred a relatively straightforward checklist, rather than an excessively intricate and burdensome one. Moreover, the checklist's interpretability is justified by its capacity to capture the perceived IC in isolation from other oral elements, such as speaking fluency and pronunciation. The checklist developers have effectively separated the IC concept from other dimensions of the speaking construct.

Raters confirmed the checklist's usefulness in interpreting and making decisions, emphasizing its practicality in everyday use.

> The interactional task is in line with both instructional task in classroom and realistic situations where one has to communicate with native speakers. That's what I appreciate

> about this checklist! I can't think of any delusional items on it. (Rater B)
>
> The good thing about this checklist is that it enables us to generalize the results to other interactional settings. What the checklist focuses on is exactly what is expected in a successful interaction. (Rater A)

This assertion can be understood in light of the nature of IC itself, positing that language is not used in isolation, but rather in conjunction with others to achieve social actions and shared life experiences. In the early stages of developing interactional skills, language users employ a limited range of linguistic tools to fulfill various social needs and situations. As IC evolves, speakers fine-tune their language to adapt to the social context, utilizing a diverse repertoire of sequential and linguistic cues accordingly (Pekarek Doehler, 2018; Roever & Kasper, 2018).

From a pragmatic standpoint, IC encompasses various pragmatic phenomena such as speech acts, discourse, implicature, politeness, pragmatic markers, NVB, and prosody. These phenomena are concerned with the interpretation of utterances within specific contexts. Interactions, which are prevalent in everyday life, are characterized by spontaneity, co-construction, and real-time emergence. Prior utterances from both the speaker and the recipient significantly shape subsequent utterances. Unlike written language which mainly focuses on grammatical correctness at the sentence level, conversational language does not solely adhere to these norms. In fact, "conversational language is not a chaotic and distorted version of written language; indeed, written language is in some senses a brittle and narrowly systematized form of spoken language, unsuitable as a model for teaching the spoken language" (Campbell-Larsen, 2015, p. 270). Considering these factors, the IC checklist items, despite conflicting with learners' conventional understanding of language learning, are likely the most valuable language skills for social situations.

Considering all these factors, it can be argued that test users can effectively interpret the scores and accompanying feedback provided by the IC checklist due to its alignment with a communicative-based curriculum, user-friendly interface, conciseness, practicality in real-world settings, and objectivity. The assumption that test users process the ability to

interpret the scores and feedback is supported by the gathered evidence. A thorough discussion of the three research questions reveals that the decision inference is to a large extent supported.

## Conclusion and Implications

In an attempt to bridge the gap between the theoretical and practical usefulness of IC in testing, the findings displayed insights about the functionality of the IC checklist in terms of decisions it brings about as well as important suggestions for the betterment of the current instrument. By developing an easy-to-use tool for IC skills, the checklist developers have contributed to promoting meaningful feedback for IC learners and integrating learning and assessment in innovative ways. The implications of these results extend to various areas of research.

The IC checklist offers candidates learning-oriented feedback on their strengths and weaknesses in IC skills during their preparation courses as well as a peer- and self-assessment instrument supporting their IC development. The checklist has the potential to guide learners toward a better understanding of the IC construct which conceivably leads to their interactional strategy training, exam preparation, and study habits. Beyond the classroom context, the authenticity of the checklist items suggests its practical applicability in everyday conversations, where individuals draw from their interactional repertoire to engage in meaningful oral discourse.

For teachers, the checklist provides a suitable means of offering feedback with a focus on learning outcomes. It also has a significant impact on instructional practices, allowing teachers to align their lesson plans, syllabi, in-class activities, and teaching techniques with the targeted IC skills. The findings can contribute to increasing raters' awareness of macro- and micro-IC features during the standardization, rating, and norming programs, enhancing their confidence in assigning IC scores. Finally, decision-makers can benefit from this research by promoting fair test practices in terms of minimizing bias, adopting systematic and collective actions, and providing timely and clear score reporting to candidates.

A notable limitation of the current investigation pertains to the potential for task-specific variations in the salience and relevance of the IC checklist items. As the IC construct is inherently context-sensitive and task-specific (May et al., 2020), the discussion tasks employed in this study may have differentially activated certain competencies over others. For instance, checklist features related to "negotiation towards an outcome" may have been less prominently displayed. This contextual constraint is an important consideration, as it suggests that the manifestation of IC can be shaped by the specific task demands and interactional requirements placed on the participants.

Here, at the end of all things, the domain of IC in general, and the assessment of the construct using the IC checklist and all its associated testing conceptualizations in particular, are still largely understudied. As Young (1992, p. 120) aptly stated regarding the IC construct over 30 years ago: "There is, thank goodness, much work still to be done."

## Author disclosures

The authors declare that they have no conflict of interest. The authors have no personal, academic, professional, or commercial relationships that might be seen to influence the conduct of the study or the conclusions and recommendations presented.

Both authors, Zahra Montasseri and Alireza Ahmadi, contributed equally to all aspects of this work, including conceptualization, design of methodology, provision of research resources, data collection, data analysis, writing the manuscript, and review and editing of the article.

## ORCID iDs

Zahra Montasseri https://orcid.org/0000-0003-3988-482X

Alireza Ahmadi https://orcid.org/0000-0001-8327-2420

## References

Banerjee, J., & Plough, I. (2016). *Behavior in speaking tests: A preliminary model of interaction* [Paper presentation]. Language Testing Research Colloquium (LTRC), Palermo, Italy.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *QUAL research in psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. Virginia: Ascd.

Cambridge Assessment English (2024, June 25). B2 First exam format. *Cambridge English*. Retrieved from https://www.cambridgeenglish.org/exams-and-tests/first/exam-format/

Campbell–Larsen, J. (2015). Interactional competence in second language acquisition. *Kwansei Gakuin Univ Humanit Rev*, *19*, 265–86.

Chalhoub–Deville, M. (2003). *Second language interaction: Current perspectives and future trends. Language Testing, 20*(4), 369–383. https://doi.org/10.1191/0265532203lt264oa

Chalhoub–Deville, M., & Deville, C. (2005). A look back at and forward to what language testers measure. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 815–832). Mahwah, NJ: Lawrence Erlbaum Associates.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement.* Peter Lang.

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualizations, operationalizations, and outstanding questions. *Language Assessment Quarterly*, *15*(3), 219–236. https://doi.org/10.1080/15434303.2018.1453816

Green, R. (2013). *Statistical analyses for language testers.* U.K.: Palgrave McMilan.

Gounaili, K. (2011). *Focusing on eye contact: Interpersonal communication among students at Eastern Mediterranean University* [Unpublished doctoral dissertation, Eastern Mediterranean University (EMU)].

Hamilton, J., Reddel, S., & Spratt, M. (2001). Teachers' perception of online rater

training and monitoring. *System, 29*, 505–20. https://doi.org/10.1016/S0346-251X(01)00036-7

Jenkins, S., & Parra, I. (2003). Multiple layers of meaning in an oral proficiency test: The complementary roles of nonverbal, paralinguistic, and verbal behaviors in assessment decisions. *The Modern Language Journal, 87*(1), 90–107. https://doi.org/10.1111/1540-4781.00180

Kley, K. (2015*). Interactional competence in paired speaking tests: Role of paired task and test-taker speaking ability in co-constructed discourse* [Unpublished doctoral dissertation]. The University of Iowa.

Knoch, U., & Chapelle, C. A. (2017). Validation of rating processes within an argument-based framework. *Language Testing, 35*(4), 477–499. https://doi.org/10.1177/0265532217710049

Konzett-Firth, C. (2020). Co-adaptation processes in plenary teacher-student talk and the development of L2 interactional competence. *Classroom Discourse, 11*(3), 209–228. https://doi.org/10.1080/19463014.2019.1597744

Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. Longman.

Lam. D., Galaczi, E., Nakatsuhara, F., & May, L. (2023). Assessing interactional competence: exploring ratability challenges. *Applied Pragmatics, 5*(2), 208-233.

Lieberman, R. J., & Michael, A. (1986). Content relevance and content coverage in tests of grammatical ability. *Journal of Speech and Hearing Disorders*, *51*(1), 71–81. https://doi.org/10.1044/jshd.5101.71

Linacre, J. M. (2021). Rasch-Model Computer Programs. Winsteps. https://www.winsteps.com/a/Facets-Manual.pdf

Liubashenko, O., & Kavytska, T. (2020). Strategy to assess L2 interactional competence of university students: Ukrainian context. *Changing language assessment: New dimensions, new challenges, 5*(1),253-274. https://doi.org/10.1007/978-3-030-42269-1_11

May, L. (2009). Co–constructed interaction in a paired speaking test: The rater's

perspective. *Language Testing, 26*(3), 397–421. doi:10.1177/0265532209104668

May, L., Nakatsuhara, F., Lam, D., & Galaczi, E. (2020). Developing tools for learning oriented assessment of interactional competence: Bridging theory and practice. *Language Testing, 37*(2), 165–188. https://doi.org/10.1177/0265532219879044

McNamara, T. (1996). *Measuring second language proficiency.* Longman.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of applied measurement, 4*(4), 386-422.

Nakatsuhara, F. (2013). *The co–construction of conversation in group oral tests.* Frankfurt am Main, Germany: Peter Lang.

Nakatsuhara, F., May, L., Lam, D., & Galaczi, E. (2018). *Learning-oriented feedback and interactional competence.* Cambridge Assessment English (Research Notes, Issue 70).

Pekarek Doehler (2018). Elaborations on L2 interactional competence: the development of L2 grammar-for-interaction. *Classroom Discourse, 9*(1), 3-24, 10.1080/19463014.2018.1437759

Purpura, J. E., & Turner, C. E. (2014). A learning-oriented assessment approach to understanding the complexities of classroom-based language assessment. *Roundtable on learning-oriented Assessment in Language Classrooms and Large-Scale Contexts*. Teachers College: Columbia University, New York.

Ramazani, M., Behnam, B., & Ahangari, S. (2019). Psychometric characteristics of a rating scale for assessing interactional competence in paired–speaking tasks at the micro-level. *Journal of English Language Pedagogy and Practice, 11*(23), 180–206. https://dx.doi.org/10.30495/jal.2019.664545

Rauthmann, J. F., Seubert, C. T., Sachse, P., & Furtner, M. R. (2012). Eyes as windows to the soul: Gazing behavior is related to personality. *Journal of Research in Personality, 46*(2), 147-156.

Roever, C., & Kasper, G. (2018). Speaking in turns and sequences: Interactional

competence as a target construct in testing speaking. *Language testing*, *35*(3), 331-355.

Stevens, D. D., & Levi, A. J. (2005*). Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning*. United States: Stylus Publishing, LLC.

Suárez, M. D. M., & Gesa, F. (2019). Learning vocabulary with the support of sustained exposure to captioned video: Do proficiency and aptitude make a difference? *The Language Learning Journal, 47*(4), 497–517. https://www.tandfonline.com/action/showCitFormats?doi=10.1080/09571736.2019.1617768

Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing 18*(3), 275–302. https://doi.org/10.1177/026553220101800302

Taylor, L., & Wigglesworth, G. (2009). Are two heads better than one? Pair work in L2 assessment contexts. *Language Testing, 26*(3), 325–339. doi:10.1177/0265532209104665

Uludag, P., McDonough, K., & Trofimovich, P. (2022). Exploring shared and individual assessment of paired oral interactions. *Studies in Language Assessment, 11*(2), 1–24.

van Compernolle, R. A. (2013). Interactional competence and the dynamic assessment of L2 pragmatic abilities. In J.R. Ross, & G. Kasper (Eds.), *Assessing second language pragmatics* (pp. 327–353). Palgrave Macmillan.

Vázquez, V. P., Lancaster, N., & Callejas, C. B. (2020). Keys issues in developing teachers' competences for CLIL in Andalusia: training, mobility and coordination. *The Language Learning Journal, 48*(1), 81–98. https://doi.org/10.1016/j.pragma.2017.01.015

Vidaković, I., & Galaczi, E. G. (2013). The measurement of speaking ability 1913-2012. In C. J. Weir, I. Vidaković, & E. D. Galaczi (Eds.). *Measured constructs: A history of Cambridge English Language Examinations 1913–2012* (pp.257- 346).

Cambridge: UCLES/Cambridge University Press.

Vo, S. T. (2019). *Effects of task types on interactional competence in oral communication assessment* [Unpublished doctoral dissertation]. University of Iowa, Iowa, the United States.

Walsh, S. (2012). Conceptualising classroom interactional competence. *Novitas-royal (Research on Youth and Language), 6*(1), 1–12.

Whalen, C. S. (2015). *Is the reception better on a different channel? Interpersonal communication satisfaction of introverts and extraverts during face-to-face versus instant messenger conversations* (Bachelor's thesis, Claremont College). Retrieved from https://scholarship.claremont.edu/cgi/viewcontent.cgi?article=1582&context=scripps_theses

Wilson, A., & Urick, A. (2021). Cultural reproduction theory and schooling: The relationship between student capital and opportunity to learn. *American Journal of Education*, *127*(2), 193-232. https://doi.org/10.1086/712086

Young, R. F. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 426-443). Routledge.