

## Original Paper

# Operationalizing Pragmatic Difficulty: Speech Act Hierarchy and the Appropriateness Gap in an AI-Generated Business English Oral Test

Shengtao Xie<sup>1\*</sup>

<sup>1</sup> School of Foreign Languages, Beijing Forestry University, Beijing, China

\* Corresponding author: Claudem917@163.com

Received: September 03, 2026 Accepted: September 25, 2026 Online Published: October 13, 2026

doi:10.22158/eltls.v8n5p61

URL: <http://dx.doi.org/10.22158/eltls.v8n5p61>

### Abstract

*Large language models can now draft scenario-based test tasks on demand, which raises a validation question that has attracted little attention: does the difficulty of the generated tasks behave as the construct predicts? This study examined an AI-generated oral test of three business speech acts to establish whether the tasks reproduced the theoretically expected ordering of difficulty, where within the rating framework difficulty was located, and whether the ability measured was separable from general business oral proficiency. Thirty business English undergraduates responded to twelve ChatGPT-generated tasks; two trained raters independently scored five performance dimensions, and a simulated BEC oral test supplied an external criterion. Scoring reliability was high (total ICC = .953; Cronbach's  $\alpha$  = .86). A  $3 \times 5$  repeated measures analysis of variance revealed a significant effect of speech act ( $p$  = .004, partial  $\eta^2$  = .213): apologies were significantly more difficult than requests and refusals, which did not differ from each other. A significant effect of rating dimension ( $p$  < .001, partial  $\eta^2$  = .237) located the greatest weakness in pragmatic appropriateness and the least in discourse coherence, and a non-significant interaction indicated that this profile held across every act. Test totals correlated strongly but not perfectly with the external criterion ( $r$  = .778). The results suggest that construct-informed prompting can produce AI-generated pragmatic tasks with theoretically differentiated difficulty profiles, while pragmatic appropriateness emerged as a relatively stable area of weakness across speech acts.*

### Keywords

*pragmatic difficulty, speech acts, pragmatic appropriateness, AI-generated assessment, business English, construct validity*

## 1. Introduction

Every test of pragmatic ability rests on an assumption about difficulty: that the situations it presents are not equally demanding, and that the differences among them follow from the construct rather than from the wording of individual items. For speech acts, the assumption is well grounded in theory. Asking a favour, turning a request down, and apologising for a mistake place different claims on the hearer's face and demand different amounts of interpersonal work (Brown & Levinson, 1987). If a test's scores rise and fall for reasons unrelated to such differences, the boundary between what learners can do and what the items happened to ask of them becomes blurred. Difficulty, on this view, is not incidental to pragmatic assessment; it is one of the properties that must be shown to behave as the construct predicts. In practice, however, difficulty in pragmatic testing is rarely engineered. Productive tasks are normally composed by hand, one scenario at a time, and the social variables that drive difficulty (relative power, social distance, the magnitude of the imposition, and the domain of the encounter) are carried implicitly in the narrative instead of being fixed by design. An item writer may intend one task to be harder than another, but nothing in the procedure ensures that an apology task is more demanding than a request task in the way the theory anticipates, or that two tasks written for the same act are comparable in load. Because each new form of a test requires fresh authorship, the resulting difficulty profile is neither transparent nor reproducible; and since expert item writing is costly, the number of tasks an operational test can contain is limited, which in turn restricts what score patterns can reveal about the construct.

Generative artificial intelligence alters both the economics and possibly the logic of this problem. When a large language model (LLM) drafts scenario-based tasks, the contextual variables of the scenario can be supplied directly in the prompt. Power, distance, imposition and domain may be set as parameters, crossed deliberately, and re-specified for a new form without new authorship (Kohnke et al., 2023; Shin & Lee, 2024). Difficulty then becomes something a developer can specify in advance rather than something to be inferred after administration. Work on automatic item generation has pursued a comparable aim for other constructs, but earlier generators depended on predefined templates and substitutions and could not readily encode the situational reasoning that pragmatic tasks require (Attali et al., 2022; Bormuth, 1969; Circi et al., 2023; Sommer & Arendasy, 2025). Whether LLM-generated pragmatic tasks now behave as the construct predicts is nonetheless an empirical question, and one that cannot be settled by inspecting the plausibility of the generated text.

Research on AI in language assessment has so far concentrated on feasibility, item quality and scoring (Lin & Chen, 2024; Mizumoto & Eguchi, 2023; O'Grady, 2023; Shin & Lee, 2024). Questions about the internal coherence of difficulty have been left largely untouched: whether, for instance, the acts that theory identifies as most face-threatening do in fact yield the lowest scores, or whether the dimension on which learners struggle most is the one that models of communicative competence single out. The question matters because difficulty that does not track the construct cannot be interpreted. A low score may indicate limited pragmatic ability, or merely an aggressive prompt; without evidence that the difficulty structure is ordered, the two explanations cannot be told apart, and neither score

comparability across tasks nor defensible task selection is possible.

This article addresses the question through an AI-generated oral test of three business speech acts: requests, refusals and apologies. It forms part of a wider investigation into AI-generated pragmatic assessment, but the analysis reported here is deliberately narrow. Rather than evaluating the instrument as a whole, it asks whether the generated tasks produce a theoretically coherent difficulty structure, where within the rating framework difficult performance is located, and whether the ability being measured can be separated from general business oral proficiency. Three questions guide the study:

RQ1. Do the AI-generated tasks reproduce the theoretically predicted ordering of difficulty across requests, refusals and apologies?

RQ2. Is difficulty localised in particular rating dimensions, and does the dimension profile remain stable across the three speech acts?

RQ3. To what extent are scores on the AI-generated pragmatic tasks associated with general business oral proficiency?

## **2. Pragmatic Difficulty: Sources, Expectations and Design**

### *2.1 Why Speech Acts Differ in Difficulty*

Speech act theory established that utterances perform actions and that any given action may be realised in many ways (Austin, 1962; Searle, 1969). What makes one realisation harder to produce than another has been answered chiefly through politeness. Brown and Levinson (1987) proposed that face-threatening acts vary in weight as a joint function of the social distance between participants, the relative power of the hearer, and the magnitude of the imposition, and that speakers select redressive strategies accordingly. Leech (1983) described politeness as a set of competing maxims, and Thomas (1983) distinguished sociopragmatic failure, which concerns the appropriateness of the social judgement, from pragmalinguistic failure, which concerns the linguistic means chosen to express it. Difficulty, on these accounts, is not a property of an act in the abstract but of an act performed under specified social conditions.

The three acts examined here are expected to differ in a consistent direction. Apologies are unusual in threatening the speaker's own positive face: the speaker must admit responsibility, express regret, and usually commit to repair, all of which carry self-presentational cost and cannot be deflected by mitigating the force of the utterance, because the speaker's own fault is the topic (Blum-Kulka et al., 1989; Brown & Levinson, 1987). Refusals threaten the hearer's positive face, yet the strategy repertoire available for them is substantial—appreciation, explanation, alternatives, deferral—and the catalogue of these options is well documented, so a learner who knows the repertoire can offset much of the threat. Requests threaten the hearer's negative face, but conventional indirectness provides routinised solutions that instruction tends to supply. Apologies should therefore sit at the difficult end of the scale, with requests and refusals some distance below.

Empirical work on task difficulty supports the general expectation while cautioning that the act is only

one input. Huang and Lu (2023) showed that the difficulty of oral speech act production tasks varies systematically with features of the task itself, and methodological reviews have long noted that the elicitation format shapes performance independently of the construct (Golato, 2003; Kasper & Dahl, 1991). Difficulty is thus a joint product of the act and the conditions under which it is elicited. The prediction tested here is correspondingly constrained: within a set of tasks that hold the social parameters parallel across acts, scores should still order by act, with apologies lowest.

### *2.2 The Dimension that Lags Behind*

Models of communicative competence separate grammatical, sociolinguistic, discourse and strategic knowledge rather than treating proficiency as a single quantity (Bachman, 1990; Canale & Swain, 1980), and this separation underpins the multi-dimensional rating of pragmatic performance (Roever, 2006; Timpe-Laughlin et al., 2015). A body of evidence indicates that the components do not develop in step. Bardovi-Harlig and Dörnyei (1998) found that instructed learners detected grammatical violations more readily than pragmatic ones, and comparable asymmetries have been reported for learners whose linguistic resources are advanced while their contextual judgement remains insecure (Liu, 2006; Taguchi, 2015). The pattern is often summarised as a lag: accuracy and fluency are consolidated before appropriateness becomes reliable.

The lag has a direct implication for scored performance. Where several dimensions are rated from the same response, the dimension that depends on contextual judgement should be the least secure and therefore the lowest-scoring, while dimensions tied to linguistic form should be comparatively protected. This yields a testable expectation that is independent of the particular speech act: if appropriateness is genuinely the lagging component rather than an artefact of one task type, then it should be the weakest dimension for requests, refusals and apologies alike. Demonstrating that stability would also have a practical consequence, since a dimension profile that holds across tasks licenses comparable scoring across them.

### *2.3 Difficulty as a Design Parameter*

If difficulty is a function of specifiable social variables, it should in principle be designable. Automatic item generation has long been motivated by exactly that prospect, promising items calibrated to a construct model rather than composed ad hoc (Bormuth, 1969). The tradition encountered a familiar obstacle: systems built on rules and templates varied surface features reliably but had no mechanism for the contextual reasoning that pragmatic items demand, so the difficulty they produced was shallow and predictable, while model-based generators were only as good as the construct specification they were given (Attali et al., 2022; Circi et al., 2023; Sommer & Arendasy, 2025).

Standard test development does not leave difficulty to chance for long. Items are trialled, their difficulty estimated from response data, and forms assembled so that the demands placed on candidates are known rather than assumed (Bachman, 1990; Weir, 2005). Pragmatic assessment has struggled to follow this routine. Because producing one task requires an elaborate situation and a judgement about its social configuration, item pools remain small, and a pool of a dozen or so tasks cannot support

stable item-level estimates; difficulty is therefore usually addressed qualitatively, through the item writer's intentions or the test's face plausibility, rather than demonstrated empirically. The consequence is a gap between what a pragmatic test claims to measure, namely the ability to act appropriately under specifiable social conditions, and what it can show about the conditions it actually creates. Closing that gap requires tasks whose social parameters are known by construction, which is precisely what parameterised generation can supply.

LLMs remove part of that obstacle because a prompt can state the construct variables in words. Power, distance, imposition and domain can be enumerated, crossed and recombined, and the same prompt template can be reused across forms, which makes the difficulty of the resulting task set in principle auditable before administration. Early applications in language testing have produced encouraging evidence for reading and vocabulary items, essay scoring, and conversational material for pragmatic analysis, and studies of chatbot interaction have begun to examine pragmatic development directly (Chen et al., 2024; Kohnke et al., 2023; Lin & Chen, 2024; Mizumoto & Eguchi, 2023; Rahman et al., 2025). What remains untested is whether the difficulty that such prompts parameterise is the difficulty the construct implies. That is the gap this study addresses, using an instrument whose tasks were generated from an explicit matrix of social variables.

### **3. Method**

#### *3.1 Design and Scope*

The study reports a quantitative validation of the difficulty structure of an AI-generated oral pragmatic test for business English. The data derive from a larger project on AI-generated pragmatic assessment; the variables analysed here are limited to task-level performance, the five rating dimensions, and the external criterion, because these are the sources of evidence that bear on the difficulty questions posed above. Other strands of the project, which concern how learners perceive the authenticity of generated scenarios and what consequences testing has for their subsequent learning, are reported separately and are not duplicated here.

#### *3.2 Participants*

Thirty undergraduate students majoring in business English at a Chinese university took part. All had completed at least two years of advanced English instruction and were at intermediate to upper-intermediate level, as indicated by their having passed the Test for English Majors Grade Four. None had received formal instruction in interlanguage pragmatics, so their performance could be taken to reflect naturally developed pragmatic ability rather than trained responses. Participants were recruited voluntarily and were informed that results would not affect their academic records. They varied in exposure to workplace communication and in overseas experience, which is relevant because the tasks assume familiarity with business roles.

Because previous instruction had concentrated on grammar and lexis, the sample provided a reasonably stable baseline in linguistic terms while leaving pragmatic strategy use free to vary, that the condition

under which differences in scored performance can be attributed to contextual judgement rather than to uneven linguistic resources.

### 3.3 Task Generation

Twelve oral tasks were produced by ChatGPT-4o from prompts designed for the study, with four tasks for each act. Every prompt specified the same set of construct variables so that the three acts could be compared: the role relationship and its power configuration (for example student above professor, intern below manager, colleague at the same level), social distance (low, moderate or high), the magnitude of the imposition (low, medium or high), and the domain of the encounter (academic work, internship, business meeting, negotiation, or client relations). Each resulting task contained a short situation establishing the roles, the contextual constraints and the goal, and asked the candidate for a single spoken response.

To ensure adequate coverage of the three target speech acts, the test included twelve oral tasks, with four tasks for each speech-act category: requests, refusals, and apologies. The request tasks required participants to make polite requests in four different situations: asking a professor for a two-day extension to submit an individual reflection report, asking a senior representative from a partner company for a one-day extension to finalize a pricing document, asking a teammate for help with data charts, and asking a supervisor for assistance in obtaining information from another department. The refusal tasks involved four situations with different interpersonal and institutional constraints: refusing a colleague's request to take over a client-related task, declining an additional discount requested by a long-term client, refusing a classmate's request to copy part of an assignment, and declining a senior team member's request to complete an additional proposal under a conflicting deadline. The apology tasks likewise covered four situations involving responsibility for different types of problems: failing to complete a customer-feedback summary on time, sending incomplete product information to an important client, failing to upload the final version of presentation slides, and presenting outdated sales data during a team meeting. Across the twelve tasks, the scenarios varied in terms of role relationship, institutional setting, degree of imposition, and communicative purpose. This distribution reduced the likelihood that an observed speech-act difference would be attributable to a single scenario type while retaining a common focus on business and academic communication.

Prompt design proceeded in two rounds. The first round used open instructions, for example a request to generate a polite request scenario between a student and a professor. Inspection of the output showed that the social variables were encoded unevenly and that lexical complexity fluctuated, so difficulty was being determined by the generated wording rather than by the construct. The second round stated the variables explicitly and added constraints requiring realistic and professionally appropriate scenarios, a positive and respectful tone, cultural sensitivity, and the exclusion of sensitive content such as reference to race, gender, religion, politics, stereotypes, personal trauma or illegal activity, as well as of extreme conflict. The final template was applied uniformly to all twelve tasks.

The task set was piloted with eighteen learners drawn from the same population before the main

administration. The pilot checked whether each task elicited the intended act rather than a neighbouring one, whether the power, distance and imposition relations were recoverable from the situation description, whether lexical and syntactic load was appropriate, whether the preparation and response windows were workable, and whether any scenario rested on culturally inappropriate assumptions. Tasks that failed these checks were revised, and the timing was fixed at 15 seconds of preparation and 45 seconds of response per item.

### *3.4 Procedure and Scoring*

Testing was conducted online under standardised conditions. Each participant was paired with a trained interlocutor who delivered the prompts in a fixed order, and all responses were audio-recorded; a full session lasted approximately twelve to fifteen minutes. Before the pragmatic tasks, participants completed a simulated Cambridge Business English Certificate (BEC) oral examination, scored on the official four-dimension scale covering grammar and vocabulary, discourse management, pronunciation and interactive communication. The BEC served as an external criterion of general business oral proficiency; it does not assess pragmatic competence directly, but effective performance on it depends on contextually appropriate language use.

Two raters with expertise in language assessment and interlanguage pragmatics independently scored all recordings. Before formal scoring, the raters completed a calibration session using benchmark responses to establish a shared interpretation of the rating criteria. Each response received a score from 1 to 5 on five dimensions: pragmatic appropriateness, linguistic accuracy, fluency, discourse coherence, and overall communicative effectiveness. Inter-rater reliability was calculated using the original ratings assigned independently by the two raters, before any discussion or reconciliation of disagreements. After the reliability analysis, discrepant ratings were discussed and resolved by consensus. The resulting consensus scores were then used in the subsequent analyses.

### *3.5 Analysis*

For each participant, scores were first averaged across the four tasks within each speech-act category for each rating dimension. This produced one participant-level mean score for each combination of speech act and rating dimension. These scores were then entered into a  $3 \times 5$  two-way repeated-measures ANOVA, with speech act (request, refusal, apology) and rating dimension (appropriateness, linguistic accuracy, fluency, discourse coherence, overall communicative effectiveness) treated as within-participant factors.

Inter-rater reliability was assessed using a two-way mixed-effects, absolute-agreement intraclass correlation coefficient (ICC), based on the raters' independent pre-consensus scores. This specification was selected because the same two raters evaluated all responses and because agreement, rather than consistency alone, was of interest. Internal consistency was estimated using Cronbach's alpha for the full set of tasks and separately for each speech-act category. The consensus scores obtained after the reliability analysis were used for the substantive analyses reported below.

## 4. Results

### 4.1 Scoring Reliability

Ratings were consistent enough to support the analyses that follow. As Table 1 shows, agreement was high for the test as a whole, and every coefficient exceeded the 0.75 level conventionally regarded as acceptable in applied linguistics. Agreement was strongest for requests and weakest, though still comfortable, for apologies, and among the dimensions the highest coefficient was recorded for overall communicative effectiveness and the lowest for linguistic accuracy. Internal consistency was also adequate, with the full task set reaching .86 and the act-level subscales ranging from .75 to .80; the comparatively lower figure for refusals is consistent with the wider range of strategies that refusal tasks elicit, which introduces legitimate variation in responses.

**Table 1. Scoring Reliability of the AI-generated Pragmatic test**

| Unit of analysis      | Tasks | ICC [95% CI]      | Cronbach's $\alpha$ |
|-----------------------|-------|-------------------|---------------------|
| All tasks             | 12    | .953 [.901, .978] | .86                 |
| Requests              | 4     | .968 [.932, .985] | .78                 |
| Refusals              | 4     | .943 [.882, .972] | .75                 |
| Apologies             | 4     | .862 [.728, .931] | .80                 |
| Appropriateness       | —     | .819 [.651, .909] |                     |
| Linguistic accuracy   | —     | .802 [.620, .900] |                     |
| Fluency               | —     | .842 [.696, .921] |                     |
| Discourse coherence   | —     | .825 [.663, .912] |                     |
| Overall effectiveness | —     | .871 [.747, .935] |                     |

### 4.2 The Difficulty Ordering of the Three Speech Acts

Sphericity was violated for the effect of speech act (approximate  $\chi^2(2) = 15.478$ ,  $p < .001$ ), so Greenhouse–Geisser corrected degrees of freedom are reported. As Table 2 shows, the effect of speech act was significant with a medium-to-large effect size, indicating that the three acts were not equally demanding. Bonferroni comparisons placed both requests ( $M = 3.827$ , 95% CI [3.543, 4.111]) and refusals ( $M = 3.827$ , 95% CI [3.567, 4.087]) reliably above apologies ( $M = 3.660$ , 95% CI [3.432, 3.888];  $p < .05$ ), while requests and refusals did not differ from one another. The predicted ordering was therefore reproduced at the level of the act, as summarised in Table 3, although the absolute margin between the hardest and the easier acts was modest, amounting to about one sixth of a scale point.

**Table 2. Two-way Repeated Measures Analysis of Variance for Speech Act and Rating Dimension**

| Source                        | SS      | df      | MS     | F     | p     | Partial $\eta^2$ |
|-------------------------------|---------|---------|--------|-------|-------|------------------|
| Speech act                    | 2.778   | 1.404   | 1.979  | 7.829 | .004  | .213             |
| Error (speech act)            | 10.289  | 40.716  | .253   |       |       |                  |
| Rating dimension              | 35.258  | 2.501   | 14.095 | 9.005 | <.001 | .237             |
| Error (dimension)             | 113.542 | 72.529  | 1.565  |       |       |                  |
| Speech act $\times$ dimension | .622    | 4.315   | .144   | .547  | .715  | .019             |
| Error (interaction)           | 32.978  | 125.135 | .264   |       |       |                  |

**Table 3. Predicted and Observed Difficulty of the Three Speech Acts**

| Speech act | Predicted rank | M     | 95% CI         | Pairwise outcome              |
|------------|----------------|-------|----------------|-------------------------------|
| Requests   | Intermediate   | 3.827 | [3.543, 4.111] | Above apologies ( $p < .05$ ) |
| Refusals   | Intermediate   | 3.827 | [3.567, 4.087] | Above apologies ( $p < .05$ ) |
| Apologies  | Most difficult | 3.660 | [3.432, 3.888] | Lowest of the three acts      |

#### 4.3 Where the Difficulty Lies: The Dimension Profile

Sphericity was also violated for the effect of rating dimension (approximate  $\chi^2(9) = 41.861$ ,  $p < .001$ ) and for the interaction (approximate  $\chi^2(35) = 117.200$ ,  $p < .001$ ). The dimension effect was significant and again medium-to-large in size (Table 2), showing that the five dimensions were not equally attainable. Bonferroni tests identified discourse coherence as the highest-scoring dimension ( $M = 4.156$ , 95% CI [3.871, 4.440]) and pragmatic appropriateness as the lowest ( $M = 3.322$ , 95% CI [2.971, 3.673]); appropriateness fell significantly below every other dimension except fluency. The interaction between speech act and dimension was not significant, and its effect size was negligible. Substantively, this means that the ranking of the five dimensions was the same for requests, refusals and apologies: whatever the task asked of the candidate, appropriateness remained the dimension on which performance was weakest, and coherence the dimension on which it was strongest.

#### 4.4 Separation from General Oral Proficiency

Pragmatic total scores showed a strong positive correlation with scores on the simulated BEC-style oral assessment ( $r = .778$ ,  $p < .001$ , two-tailed). The coefficient indicates substantial overlap between performance on the AI-generated pragmatic tasks and general business oral proficiency. The squared correlation suggests that approximately 60% of the variance was shared between the two measures in this sample. The remaining variance should not be interpreted as a direct measure of pragmatic ability, because it may reflect task-specific variance, measurement error, or other sources of variance. The correlation therefore provides evidence of association with general oral proficiency, but it does not by itself establish discriminant validity or demonstrate that the pragmatic test measures a distinct

construct.

## 5. Discussion

### 5.1 *A Difficulty Structure that Follows the Construct*

The first question asked whether tasks generated from an explicit matrix of social variables would order themselves by speech act. They did. Apologies proved significantly harder than requests and refusals, and the latter two were indistinguishable, which is what a face-theoretic account anticipates (Brown & Levinson, 1987). Because the task set was generated using an explicit specification of role relationship, social distance, imposition, and domain, the three speech-act categories were designed to incorporate comparable contextual considerations. However, these variables were not independently manipulated or fully matched across all tasks. The observed ordering should therefore be interpreted as evidence that the speech-act differences emerged within a systematically specified task set, rather than as evidence of a fully controlled experimental effect of speech act.

The result also clarifies what the word difficulty is doing here. In psychometrics, difficulty names the proportion of candidates who succeed on an item, a purely empirical quantity. In the present argument, difficulty names a demand that the construct attributes to a situation: the amount of interpretive and interpersonal work an act requires under given social conditions. The two are related but not identical, and the modest size of the observed gap belongs to the second sense. Generated tasks can carry a theoretically heavier demand and still leave most candidates performing competently, which is what happened with apologies; whether the resulting score differences are large enough to be useful is a separate question from whether they are ordered as predicted. Conflating the two senses is what leads developers to assume that a task designed to be hard will necessarily function as hard.

Two qualifications temper the claim. The first concerns magnitude. The difference between the hardest and the easier acts was statistically reliable and moderate in effect size, but small in scale terms; for a test intended to discriminate finely among learners, a difficulty spread of a fraction of a point provides limited leverage, and it suggests the task set occupied a comparatively narrow band of difficulty. Widening the contrast in imposition, or extending the acts covered, would test the parameterisation more severely. The second qualification concerns the equivalence of requests and refusals. The absence of a difference is theoretically informative rather than merely null: refusals threaten the hearer's face, but the mitigation repertoire available to a learner—appreciation, explanation, alternatives—appears to compensate for the threat, just as conventional indirectness compensates for the imposition carried by requests (Blum-Kulka et al., 1989). Difficulty, in other words, is not a monotonic function of how threatening an act is; it depends on how much routinised support the learner can draw on when performing it. That nuance matters for item design, since it implies that difficulty is better controlled through the social variables and the availability of conventional strategies than through the choice of act alone.

### 5.2 Appropriateness as the Persistent Constraint

The second question asked whether difficulty is concentrated in particular dimensions and whether the pattern is stable across acts. Appropriateness was the weakest dimension and coherence the strongest, and the non-significant interaction showed that this profile held for all three acts. The result is consistent with the lag documented in the acquisition literature, in which learners detect and produce formally correct language more reliably than they recognise contextual inappropriateness (Bardovi-Harlig & Dörnyei, 1998; Liu, 2006; Taguchi, 2015), and with the multi-componential view of communicative competence that motivates dimension-based rating (Bachman, 1990; Canale & Swain, 1980; Timpe-Laughlin et al., 2015).

Why appropriateness should lag is not difficult to account for. Producing an utterance appropriately requires two things at once: retrieving the linguistic resources that express the intended meaning, and monitoring the social reading that those resources will receive from a particular hearer in a particular situation. In classroom settings the first is rewarded and corrected constantly, whereas the second is rarely made the object of explicit attention, and learners accordingly develop the habit of treating formal accuracy as the criterion of success (Thomas, 1983; Liu, 2006). The asymmetry observed here is therefore best understood as an artefact of practice rather than of aptitude—which is encouraging, because habits of attention are more tractable than deficits in ability.

Two consequences follow. First, the stability of the profile is a measurement asset as well as a substantive finding: if the ordering of dimensions is the same across tasks, then dimension scores can be compared across tasks without adjusting for what was asked, and a single set of reporting weights is defensible. Second, and more pointedly, an aggregate score would conceal precisely the information that matters. A learner whose totals are respectable but whose appropriateness ratings are depressed has a specific and remediable problem that a single number will not reveal. Reporting dimension profiles, and building instruction around the weakest one, follows directly from the evidence. Because appropriateness was the bottleneck irrespective of act, targeted work on it does not depend on selecting a particular set of scenarios: the constraint is a property of the learners' repertoire rather than of any one task type.

### 5.3 Implications for AI-generated Assessment

The third question concerned the relationship between performance on the AI-generated pragmatic tasks and general business oral proficiency. The strong positive correlation with the simulated BEC-style assessment indicates substantial overlap between the two forms of performance. This overlap is theoretically plausible because pragmatic performance necessarily draws on broader oral-language resources. At the same time, the correlation was below unity, indicating that the two measures were not identical in this sample. However, the unshared variance cannot be attributed specifically to pragmatic ability on the basis of a bivariate correlation alone. Additional validity evidence, such as correlations with established measures of pragmatic competence, comparisons with theoretically distinct constructs, or a multivariate validity design, would be needed to evaluate the

distinctiveness of the pragmatic construct more directly.

Taken together with the difficulty findings, this supports a specific role for LLM-generated tasks. Their value lies less in replacing expert item writing wholesale than in making difficulty an explicit, inspectable part of test construction. Three practical implications follow. Construct variables should be written into prompts rather than left to the model's discretion, since the first generation round in this study showed that open instructions allow difficulty to drift with wording. Generated sets should be piloted for elicitation and load before use, as the pilot in this study did. And generated tasks should be subjected to an empirical difficulty check, because the present results show that theoretical coherence has to be demonstrated rather than assumed. Where such checks are routine, generation could also supply the large numbers of comparable tasks that low-stakes formative assessment requires, and could be paired with automated scoring and feedback as those capabilities mature (Mizumoto & Eguchi, 2023; Rahman et al., 2025; Shin & Lee, 2024).

#### *5.4 Limitations*

Four limitations bound the conclusions. The sample comprised thirty learners at one institution, and although this is adequate for estimating the effects reported, it limits generalisation to other populations and prevents stable estimation of item-level parameters. Only three speech acts were examined, so the ordering established here cannot be extended to acts such as complaints, suggestions or compliments without further study. The task format elicited monologue responses, which excludes the negotiation of meaning and the repair sequences that characterise authentic interaction and have long been identified as a constraint on production tests (Golato, 2003; Kasper & Dahl, 1991); difficulty may behave differently when an interlocutor reacts. Finally, scoring remained the work of human raters, so nothing can be concluded here about automated pragmatic scoring, and because testing occurred on a single occasion the stability of the difficulty ordering over time is unknown. The scope of the article is also deliberately partial: questions about how learners perceive the authenticity of generated scenarios, and about the learning consequences of such testing, fall outside the present analysis and are addressed elsewhere in the wider project.

### **6. Conclusion**

This study asked whether the difficulty of AI-generated pragmatic tasks behaves as the construct predicts. For an oral test of three business speech acts generated from an explicit matrix of social variables, the answer was affirmative in structure and qualified in degree. Scoring reliability was high enough to make the analysis meaningful; apologies were significantly more difficult than requests and refusals, which were equivalent; pragmatic appropriateness was the weakest of five rated dimensions while discourse coherence was the strongest, and that profile held across every act; and test totals correlated strongly but imperfectly with an external measure of general business oral proficiency, leaving a substantial share of variance unshared.

The contribution is twofold. Substantively, the study supplies direct evidence that the difficulty

hierarchy implied by politeness theory survives into the performance of generated tasks, and that the constraint on business English oral performance lies in appropriateness rather than in any particular speech act. Methodologically, it demonstrates an approach in which difficulty is specified before generation and then verified empirically, converting a property that pragmatic tests usually inherit by accident into one that can be designed and audited. For practitioners, the practical reading is straightforward: use construct variables to steer generation, verify the difficulty of the result instead of presuming it, and report dimension profiles so that the weakest dimension remains visible rather than being averaged away.

Extension of this work would require larger and more varied samples, a wider range of speech acts, and interactive formats in which a partner responds; it would also need to establish that the difficulty ordering is stable across occasions and to test whether automated scoring can recover the dimension profile reported here. Each of these would sharpen the claim that pragmatic difficulty is a property that can be operationalised deliberately, rather than a by-product of who wrote the items.

## References

- Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & Von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5, 903077. <https://doi.org/10.3389/frai.2022.903077>
- Austin, J. L. (1962). *How to do things with words*. Oxford University Press.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.
- Bardovi-Harlig, K., & Dörnyei, Z. (1998). Do language learners recognize pragmatic violations? Pragmatic versus grammatical awareness in instructed L2 learning. *TESOL Quarterly*, 32(2), 233-259.
- Blum-Kulka, S., House, J., & Kasper, G. (1989). Investigating cross-cultural pragmatics: An introductory overview. In S. Blum-Kulka, J. House, & G. Kasper (Eds.), *Cross-cultural pragmatics: Requests and apologies* (pp. 1-34). Ablex.
- Bormuth, J. (1969). *On a theory of achievement test items*. University of Chicago Press.
- Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge University Press.
- Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1, 1-47.
- Chen, X., Li, J., & Ye, Y. (2024). A feasibility study for the application of AI-generated conversations in pragmatic analysis. *Journal of Pragmatics*, 223, 14-30. <https://doi.org/10.1016/j.pragma.2024.01.003>
- Circi, R. C., Hicks, J., & Sikali, E. S. (2023). Automatic item generation: Foundations and machine learning-based approaches for assessments. *Frontiers in Education*, 8, 858273.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum

*Associates.*

- Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage.
- Golato, A. (2003). Studying compliment responses: A comparison of DCTs and recordings of naturally occurring talk. *Applied Linguistics*, 24(1), 90-121.
- Huang, W., & Lu, X. (2023). The difficulty of oral speech act production tasks in second language pragmatics testing. *Frontiers in Psychology*, 14, 1096399. <https://doi.org/10.3389/fpsyg.2023.1096399>
- Kasper, G., & Dahl, M. (1991). Research methods in interlanguage pragmatics. *Studies in Second Language Acquisition*, 13(2), 215-247.
- Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language teaching and learning. *RELIC Journal*, 54(2), 537-550.
- Leech, G. N. (1983). *Principles of pragmatics*. Longman.
- Lin, Z., & Chen, H. (2024). Investigating the capability of ChatGPT for generating multiple-choice reading comprehension items. *System*, 123, 103344. <https://doi.org/10.1016/j.system.2024.103344>
- Liu, J. (2006). Assessing EFL learners' interlanguage pragmatic knowledge: Implications for testers and teachers. *Reflections on English Language Teaching*, 5(1), 1-22.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances. *American Psychologist*, 50(9), 741-749.
- Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2), 100050.
- O'Grady, S. (2023). An AI generated test of pragmatic competence and connected speech. *Language Teaching Research Quarterly*, 37, 188-203. <https://doi.org/10.32038/ltrq.2023.37.10>
- Rahman, G., Mudhsh, B. A., Almutairi, M., & Kouki, M. (2025). Optimizing ESL learners' speech act performance: The role of AI-powered chatbots in pragmatic competence development. *Theory and Practice in Language Studies*, 15(10). <https://doi.org/10.17507/tpls.1510.05>
- Roever, C. (2006). Validation of a web-based test of ESL pragmalinguistics. *Language Testing*, 23(2), 229-256.
- Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press.
- Shin, D., & Lee, J. H. (2024). AI-powered automated item generation for language testing. *ELT Journal*, 78(4), 446-452.
- Sommer, M., & Arendasy, M. (2025). Automatic- and transformer-based automatic item generation: A critical review. *Journal of Intelligence*, 13(8), 102. <https://doi.org/10.3390/jintelligence13080102>
- Taguchi, N. (2015). Instructed pragmatics at a glance: Where instructional studies were, are, and should be going. *Language Teaching*, 48(1), 1-50.
- Thomas, J. (1983). Cross-cultural pragmatic failure. *Applied Linguistics*, 4(2), 91-112.
- Timpe-Laughlin, V., Wain, J., & Schmidgall, J. (2015). Defining and operationalizing the construct of pragmatic competence: Review and recommendations (ETS Research Report No. RR-15-06).

*Educational Testing Service*. <https://doi.org/10.1002/ets2.12053>

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan.

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods.

*Language Testing*, 32(2), 199-225. <https://doi.org/10.1177/0265532214557113>