Studies in Applied Linguistics & TESOL at Teachers College, Columbia University, Vol. 19, No. 1, pp. 42–71
The Forum

# A Meaning-Based Multiple-Choice Test of Pragmatic Knowledge: Does It Work?

**Jorge Beltrán**
*Teachers College, Columbia University*

## INTRODUCTION

One of the most recent developments in language assessment, the assessment of pragmatic ability, continues to face a number of challenges. Among these, a key issue is that pragmatic ability remains as one of the least understood constructs in the field. While pragmatic ability has been the focus of much research in the last few decades, there is still a lack of consensus as to what components constitute this construct (Timpe-Laughlin, Wain, & Schmidgall, 2015). In addition, researchers differ in their perceptions of how it can be effectively and reliably elicited and assessed, particularly where practicality considerations may be of paramount concern. As such, along with the issue of construct definition, there is the paradox regarding which elicitation methods are to be used in this type of assessment.

Most research on pragmatic ability has been done using Discourse Completion Tasks (DCTs), particularly given the prevalence of the Speech Act Theory paradigm in the assessment of pragmatic ability. Other methods of elicitation have included multiple-choice (MC) tests, role-plays, and judgment tasks (Roever, 2011). A challenge remains, however, when trying to equate concerns of construct validity and of practicality. While role-plays provide a wide array of benefits in terms of the information being collected from participants, they are costly and require far more resources than DCTs or MC tests for their administration and scoring. The nature of the latter, nonetheless, limits the claims that can be made given test method effects (e.g., the fact that MC tests are an indirect measure that provides predesigned options as opposed to the elicitation of naturalistic responses). More research is needed in order to determine how the assessment of pragmatic ability can be approached in a way that satisfies institutional capacity and provides meaningful information about test-takers' abilities.

In an attempt to further explore this research agenda, the current study involved the piloting of a MC test of pragmatic knowledge control (PKC). The purpose of this paper is to evaluate the following: (a) to what extent a multiple-choice test can be aligned with a meaning-based theoretical model of pragmatic ability, (b) whether such a test can successfully identify different levels of pragmatic knowledge control, and (c) whether there is a relationship between PKC and participants' background characteristics, namely, length of stay in the US, time spent studying English, first language (L1), and educational background. A brief review of the literature is presented next in order to better understand current theoretical models of pragmatic ability. Particular attention is paid to Purpura's (2017) model, which was the theoretical grounding that was employed to develop the test of PKC that was piloted in this study.

## Pragmatic Ability: Defining the Construct

Several definitions and theoretical models of pragmatic knowledge have been developed throughout the last decades in the field of applied linguistics. One paradigm that has been adopted in multiple contexts has been the conceptualization of pragmatic ability as consisting of two separate yet related subcomponents: pragmalinguistics and sociopragmatics (Leech, 1983). As proposed by Leech (1983), pragmalinguistics refers to the aspect of pragmatics that is most concerned with language itself, that is, how linguistic elements are associated with certain illocutions (acts performed by speakers while uttering a phrase or sentence) and perlocutions (the effects an utterance has on the interlocutor). In contrast, sociopragmatics refers to the "sociological" aspect of pragmatics, that is, it is concerned with the conventions that make language use "socially acceptable," considering expectations related to roles, situations, and topics within a given community.

The dichotomy of pragmalinguistics and sociopragmatics, along with Speech Act Theory, have been predominant in the field of second language pragmatics assessment, and these two concepts often become the operationalized construct in many studies of pragmatic ability, even though most studies focus on either one of these concepts (Roever, 2014). Timpe-Laughlin, Wain, and Schmidgall (2015) conducted an extensive review of the literature in their attempt to develop a comprehensive theoretical model of pragmatic competence, which included an interactional component that had not previously been included. After surveying numerous models and research projects, they determined that three principles are "necessary conditions for felicitous pragmatic behavior:" meaning, interaction, and context. In their discussion, the authors highlighted the importance of recognizing that the interpretation and conveyance of meaning-in-context is the core of pragmatic ability. In doing so, they pointed out Purpura's (2004) distinction of literal, intended, and pragmatic meanings, where the latter are primarily derived from context. In this way, the model of pragmatic competence proposed by Timpe-Laughlin, Wain, and Schmidgall (2015) included four subcomponents: discourse knowledge, grammatical knowledge, pragmatic-functional knowledge, and sociocultural knowledge. These subcomponents are depicted within an interactive process between two interlocutors, where "the context mediates meaning" (p. 16), both for input and output processes.

While Timpe-Laughlin, Wain, and Schmidgall's (2015) model does provide a comprehensive picture of the types of knowledge that comprise and interact with pragmatic competence, their componential model seems to overlap with models of language ability itself. For example, it becomes unclear whether knowledge of grammar and discourse practices should be considered to be part of pragmatic competence and language ability at the same time. Although the model is described as being situated within a broader model of communicative language ability, such broad considerations of discourse and grammatical knowledge would make it more difficult to characterize a test construct as that of pragmatic competence alone (as opposed to communicative competence). Considering this and given Purpura's (2004) clear distinction between literal/semantic and implied meanings, it was decided that using his revisited model (Purpura, 2017) would best address the purposes of this study.

In Purpura's (2017) revisited model, more layers have been added to the different types of pragmatic meanings that are exchanged by interlocutors. Pragmatic knowledge, which is in fact a sub-component of language ability, is composed of *functional* and *implicational knowledge*. *Functional knowledge* allows speakers to understand *functional meanings*, which show intentionality and allow speakers to pursue courses of action through language (e.g., making a request). In contrast, *implicational knowledge* refers to the expression and conveyance that are dependent upon shared knowledge structures (such as social norms, conventions, and background knowledge). *Implicational meanings* have been found to be troublesome for language learners and are believed to be related to second language (L2) proficiency (e.g., high proficiency learners are able to focus more on the subtleties of context-rich exchanges; Purpura, 2017). According to Purpura, there are seven types of implicational meanings: *situational, sociolinguistic, sociocultural, psychological, literary, interactional,* and *rhetoric*. Next, each of these categories is discussed.

*Situational meanings* (formerly referred to as *contextual meanings*; Purpura, 2004) refer to meanings that derive from an understanding of the context or situation in which a given exchange is taking place. Understandings of reference, association, implicature, and figurative language are examples of how this type of pragmatic meaning is produced and interpreted (Purpura, 2017). *Sociolinguistic meanings* are those that depend on the understanding of the set of norms and expectations that apply to speakers within a particular speech community. Choice of register and directness are two examples of how *sociolinguistic meanings* are conveyed in interaction (Purpura, 2017). *Sociocultural* or *intercultural meanings* refer to those that depend on the understanding of the sets of norms and expectations that apply within and across given demographic and linguistic cultures. Use of humor and topic management within certain cultural circles are examples of how these meanings are conveyed (Purpura, 2017).

When speakers attempt to convey certain stance, whether this is an attitude, disposition, or the like, they are conveying *psychological meanings*. This category is often disregarded in taxonomies of *pragmatic meanings* (Purpura, 2017). Another category of pragmatic meanings, *rhetorical meanings*, is concerned with the expression and conveyance of meanings that are based on understandings of "textual structuring practices, genres, discourse modes, and coherence" (Purpura, 2017, p. 21). Thus, *rhetorical meanings* are concerned with appropriate language use in relation to expectations of organizational structures.

In consideration of the meanings that are produced through interactional actions, Purpura (2017) proposes the category of *interactional meanings*. These should be understood as the meanings that derive from appropriate (or inappropriate) use of "conversational structuring practices, sequencing practices, turn-taking practices, and repair practices" (Purpura, 2017, p. 21). It should be noted, however, that the interactional resources themselves are not considered pragmatic per se; instead, *pragmatic meanings* derive from compliance or deviation from conventional conversational practices. One last category, added in the updated version of Purpura's (2017) model, is that of *literary meanings*, which are "based on understandings linked to aesthetic imagination, fantasy, embellishment, exaggeration, and figures of speech" (p. 21). Since this type of meaning conveyance may require background knowledge (and may in itself be restricted by knowledge of grammar and vocabulary), it was not included in the operational theoretical model of the test that was developed for the current study.

Purpura's (2017) model seems to include the most comprehensive categorization of *pragmatic meanings*, since, rather than attempting to simplify the plethora of context-derived meanings that are found in conversation, he characterizes their differences and broadens the scope of what is to be assessed and elicited. For these reasons, this was the model that was selected to design the test of pragmatic knowledge control that is the focus of this paper, as it makes clearer how to distinguish and separately assess many different dimensions of pragmatic meaning for a fine-grained picture of test-takers' knowledge.

## Pragmatic Knowledge and Test-Taker Characteristics: Are They Related?

While it is often acknowledged that the development of pragmatic knowledge is influenced by numerous factors, determining how this relationship works empirically is a much more complex task. Among the variables of interest for this study, proficiency and length of stay have been the most studied in second language pragmatics assessment.

Language proficiency has been found to be a determining factor in how pragmatic routines are achieved by learners. For example, when it comes to directness, it has been found that lower-ability learners usually produce more direct language (e.g., when producing requests) than higher-ability learners (Félix-Brasdefer, 2007; Nguyen, 2008). Unsurprisingly, high-ability learners have been found to make use of a wider range of linguistic resources when attempting to convey *pragmatic meanings* (Blum-Kulka & Olshtain, 1986; Kobayashi & Rinnert, 2003). However, this relationship seems to be language-specific, and various phenomena have been found to depend on the target language under scrutiny (Bardovi-Harlig & Bastos, 2011). Thus, the field is still unclear about how language proficiency relates to pragmatic ability. While it would seem clear that L2 learners gain more linguistic resources as they build up their linguistic repertoire, this does not mean that high ability learners are inherently pragmatically competent. In fact, pragmatic ability has been found to explain more variance in overall language ability for high-ability test-takers and, hence, has been proposed as a better measure for separating proficient language learners (Grabowski, 2009). Given these findings, the unanswered question remains: To what extent does language proficiency explain pragmatic ability?

Another variable of interest has been *length of stay* (LOS), which has had mixed results in studies of pragmatic development. Some studies have found a positive effect of length of stay on pragmatic development (e.g. Barron, 2003; Félix-Brasdefer, 2004; Hoffman-Hicks, 1999) while others have found little to no effect (Bardovi-Harlig & Bastos, 2011; Bataller, 2010; VonCannon, 2006). Given the variability of the results, it has been categorized as an unreliable predictor of the development of pragmatic ability (Kasper & Rose, 2002). Despite this characterization, length of stay has continued to be a common study variable in L2 pragmatics, particularly in the context of study-abroad programs (Bardovi-Harlig & Bastos, 2011; Cohen & Shively, 2007; Shively, 2011). Many of these studies have looked at short-term effects on the production of certain speech acts, such as requests (Schauer, 2007), apologies (Barron, 2003), and refusals (Félix-Brasdefer, 2004). However, these studies have failed to produce convincing (or uniform) trends (Bardovi-Harlig & Bastos, 2011).

The remaining two variables of interest, educational background and native language, have also produced mixed results in terms of their effects on pragmatic ability. While language background has been shown to have an impact on L2 learners' pragmatic ability, more studies within the field of language assessment are needed. One of the studies that has looked into first language-derived differences in a multiple-choice test of pragmatic knowledge was conducted by Roever (2007), who found substantial Differential Item Functioning (DIF) in an item of formulaic implicature[1] between test-takers of European and Asian backgrounds, with Asian test-takers being at a disadvantage, likely caused by the lack of cultural knowledge that they had.

[1] The formulaic implicature item in this test is referred to as "the Pope Question." English speakers understand the question "Is the Pope Catholic?" as a rhetorical question, implying an emphatic "yes" through irony. The possible contexts of use for this question are limited, and L2 speakers may fail to understand the implied meaning of this routinized expression.

No previous studies on the possible relationship between educational background (e.g., last completed educational degree) and pragmatic ability were found. This, along with the inconclusive results for the other background variables of interest, signals a gap in the literature of second language pragmatics assessment.

## Research Questions

The following research questions were developed to guide this study:

1. How can a meaning-based model of pragmatic ability be used as the basis for the design of a pragmatic knowledge test?

2. How effectively can a MC test of pragmatic knowledge separate test-takers in terms of their ability?

3. To what extent is there a relationship between test-takers' time studying English, academic background, length of stay in the US, and L1 background with pragmatic knowledge?

# METHOD

In this section, an overview of the methodology that was followed in this study is provided.

## Research Design

This study follows a non-experimental, ex post facto design (Wiersma & Jurs, 2009), since participants received no experimental treatment and were only tested on their knowledge of pragmatic ability on one occasion. The variable under study was *L2 pragmatic knowledge control,* which was assessed through the piloting of the test. Four background variables were examined: length of stay in the US, length of English studies, educational background, and native language. This study made use of quantitative analyses to evaluate the quality of the test being piloted (Wiersma & Jurs, 2009). Participants were recruited through *purposeful sampling.* Classes from two proficiency levels at a community English program (CEP) were selected to pilot the instrument. Intermediate and advanced students were selected in order to compare the performance of these different ability levels and to determine whether the test could discriminate among different types of ability. An assumption was that a wider range of language proficiency would also represent a wider range of pragmatic ability (Wiersma & Jurs, 2009).

## Context of the Study

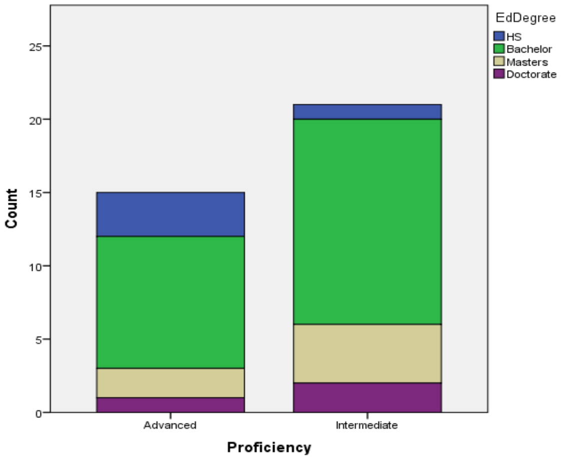
For this pilot study, data collection was conducted at an adult English as a Second Language (ESL) program at an American university in a major city. This program consists of a very diverse student population, with students from different educational backgrounds, ages, and purposes for taking ESL classes. Pragmatics is often taught explicitly in the program, but pragmatics instruction is not standardized, and it is not a main focus of assessment.

## Participants

A total of 37 participants were recruited from five different classes of the CEP, which corresponded to five different levels: Intermediate 1 (n=5), Intermediate 2 (n=10), Intermediate 3 (n=7), Advanced 1 (n=8), and Advanced 2 (n=7). Classes from the intermediate and advanced levels were chosen given one of the objectives of the project, which was determining whether the test could discriminate between high-and lower-ability test-takers. It should be noted, however, that the test was not administered to beginner test-takers, because they needed enough language ability to read the scenarios for each item in order to take the test.

The native languages of the participants were Japanese (14), Spanish (5), Chinese (6), Korean (5), Portuguese (4), French (1), German (1), and Italian (1). Participants had been in the country between six months and twenty years. Most students, however, had been in the country for two years or less (n=32). With regards to their educational level, four students had a high school degree as their highest educational degree, while the majority of participants (23) had obtained a bachelor's degree. Six students had received a master's degree, while the remaining three had earned doctorate degrees. One of the participants did not report her last completed education degree. Figure 1 shows the distribution of educational levels for the intermediate and advanced classes.

**FIGURE 1**
**Distribution of academic backgrounds by proficiency level**



## Instruments

The test that was developed to measure participants' control of pragmatic knowledge was designed following Purpura's (2017) meaning-based model of language ability. The test is comprised of 21 items. Five of these items target one of Purpura's *pragmatic meanings*, whereas the remaining 16 target two subcomponents at a time. The development of the test followed Mislevy, Steinberg, and Lukas' (2004) Evidence Centered Design (ECD), which is comprised of two components: The Conceptual Assessment Framework (CAF) and the Four-Process Delivery Architecture for Assessment. Each of these is described in detailed below.

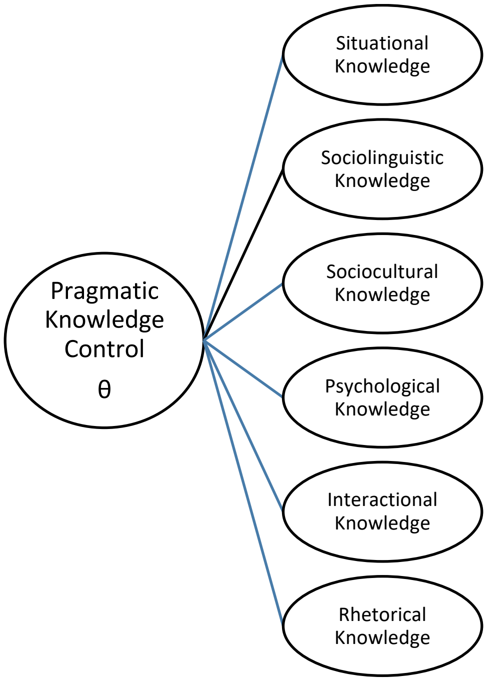
## The Conceptual Assessment Framework

Four models comprise the CAF for the test blueprint.

### *The Student Model*

For the current test, the student model consists of a single yet multi-componential unobservable variable, pragmatic knowledge control, which in turn belongs to the broader construct of language knowledge, as depicted in Purpura's (2017) model. The test specifically targets implicational knowledge, which is comprised of the types of meaning described in the literature review. Figure 2 shows the student model for this test. There is a single proficiency variable, with seven second-order factors, to be observed through the items developed in this blueprint.

**FIGURE 2**
**Proficiency model for pragmatic knowledge control (adapted from Purpura, 2017)**



### *The Evidence Model*

The evidence model is comprised by evidence rules and the measurement model.

***Evidence rules.*** In the multiple-choice test, each item is associated with one observable variable ($x_i$), and each item is to be scored dichotomously, with answers coded as correct or incorrect, after comparison with the answer key. Providing a correct response is considered as displaying evidence of control of at least one of the subcomponents of pragmatic knowledge, so that each response provides details about the pragmatic knowledge (or lack thereof) of the test-taker (the student model). In order to provide a correct response, a test-taker has to (a) assess the situation being presented, (b) evaluate the language of each possible answer, and (c) choose the answer that best fits the depicted context.

***Measurement Model.*** The measurement model used to interpret test results was Item Response Theory (IRT), specifically Rasch analysis. Dichotomous Rasch Analysis was used to analyze the relationship between the latent variable of pragmatic knowledge ($\theta$), and the test items ($x_i$). The Dichotomous Rasch Model is a probabilistic approach of mathematical modeling that allows for the measurement of persons and items on an equal-interval scale (the logit scale). It is a one-parameter model, given that "the model assumes the probability of a given person/item interaction (in terms of rating high or low) is only governed by the difficulty of the item and the ability of the person" (Granger, 2008, pp. 1122–1123). In this way, dichotomous Rasch analysis allows us to compare items and persons in term of difficulty and ability, respectively, by positioning them on the interval logit scale.

While the model works better with robust sample sizes, it has been suggested that a sample of 30 is a minimally viable sample size (Downing, 2003). Furthermore, Linacre (1994) argues that a sample of 30 is enough to obtain stable estimates in a pilot study. He also notes, however, that a small sample cannot lead to any type of "definitive statistical analysis" (Linacre, 1994, p. 328).

For future study, with a larger sample size, Structural Equation Modeling would be an ideal approach to study the second-order factors that correspond to each type of knowledge and their relationships. For a detailed diagram of the model, see Appendix C.

### *The Task Model*

In order to overcome the limitations of a selected-response task, in particular in the context of pragmatics assessment, numerous measures need to be taken to guarantee that each item provides enough context to the test-taker to allow for a single key (for the answer key and item coding scheme of sample items, see Appendix B). In order to ensure that parallel items can be generated in the future, the following task shell was generated.

### *The Assembly Model*

At this stage, the pilot test of pragmatic knowledge does not have an assembly model. It is hoped that information collected from this pilot, such as item difficulty, helps establish the relationship between the items and the task model characteristics described in the task shell above.

**TABLE 1**
**Task Shell for the Multiple-Choice Test of Pragmatic Knowledge Control**

| | | *Task Model* | |
|---|---|---|---|
| *What is being measured?* | *Fixed elements* | *Variable elements* | *List of variants for each* |
| (Receptive) Control of pragmatic knowledge in high-context situations<br>Claim: Test takers can apply pragmatic knowledge to evaluate and choose the best language to be used in the situations that are presented to them based on the context description and dialogue.<br>Measurement: Pragmatic Knowledge (operationalized as the implicational subcomponents of pragmatic knowledge described in Purpura, 2017). | 1. Nature of the task:<br>-Evaluate the situation and select the best option to appropriately complete the dialogue.<br>-Instructions.<br>-Description of the situation. All the important contextual information should be included, preferably between 40-80 words.<br>-The dialogue of item types *b* and *c* should be integrated coherently, in a way that redundancy is avoided, and there is no overload of words.<br>-The test should include at least three items per sub-component of PKC.<br>2. Order of item elements:<br>(a) Situation—clear description of the context.<br>(b) Guiding question or direction (within situation) —brief question or directive framed in relation to the situation.<br>(c) For item types *b* and *c*, dialogue—indented, formatted with interlocutors' names, and a blank for the test-taker to fill in. with a distractor.<br>(d) Options—four options, three distractors and one key, ordered from shortest to longest. | 1. Number of subcomponents of PKC being targeted.<br>2. The contextual factors that are used to elicit the various parts of PKC targeted by the test.<br>3. The topic being discussed.<br>4. The distribution of interactive patterns (priority is given to the types of pragmatic meanings covered). | 1. One or two subcomponents from Purpura's (2017) model per item.<br>2. Social distance, power, imposition, stance, conversational structure, sociocultural considerations, actions.<br>3. Various topics at the social-interpersonal, social-transactional, academic, and professional domains.<br>4. Items should fall within one of these categories:<br>(a) Turn initiation—test-taker chooses one answer to begin a conversation.<br>(b) Response—test-taker chooses the best reply to a first-pair part.<br>(c) Insert-sequence—test-taker chooses the best sentence (s) to complete a dialogue. |

## Procedures

### *Data Collection*

The test was administered on four different dates, during class time, for each of the classes. On average, the test was completed in 25 minutes.

### *Data Analysis*

***Scoring procedures.*** All items were scored dichotomously, with each item having only one correct response (key, coded as 1) and three incorrect distractors (coded as 0). The scores were added to compute a total score. According to the item codings, seven subscores for each of the different types of meanings were computed. For example, Items 8, 9, 12, and 13 targeted *psychological meanings*, so that the subscore for this type of meanings would be calculated with a maximum of four points if a test taker had answered all those items correctly. Similar procedures were followed for the other types of pragmatic meanings.

***Data analyses.*** In addition to calculating descriptive statistics, Cronbach's Alpha was calculated as the reliability indicator for the scale using the computer software program SPSS Version 25 for Windows 10 (IBM Corporation, 2017). Data for this study were analyzed using the program WINSTEPS (Linacre, 2015). Due to the small sample size for this study, Rasch analysis was used in an exploratory fashion. In addition, multiple regression analysis was conducted using the computer software program SPSS Version 25 for Windows 10 (IBM Corporation, 2017) to determine whether any of the background variables under study was a meaningful predictor of pragmatic ability (operationalized as the total score in the test) or any of its subcomponents (e.g., psychological or interactional meanings).

***Model for multiple regression analyses***. The outcome variable of interest is Pragmatic Knowledge Control [VAR: TotalPrag]. This variable was an interval scale, with scores ranging between a minimum of 0 and a maximum of 21. Multiple predictor variables were used to evaluate the relationship of these predictors with pragmatic knowledge. These variables were proficiency (operationalized as the CEP level that students were in, Intermediate and Advanced, thus becoming a categorical variable); length of stay in the US (interval variable, in months); time studying English (interval variable, in months); academic background (operationalized as their last educational degree, a categorical variable, with the four values described previously); and L1 (categorical variable, with seven different values).

# RESULTS

## Descriptive Statistics

Measures of central tendency and dispersion were computed in order to better understand how test-takers performed on the test. Table 2 shows the descriptive statistics for the test.

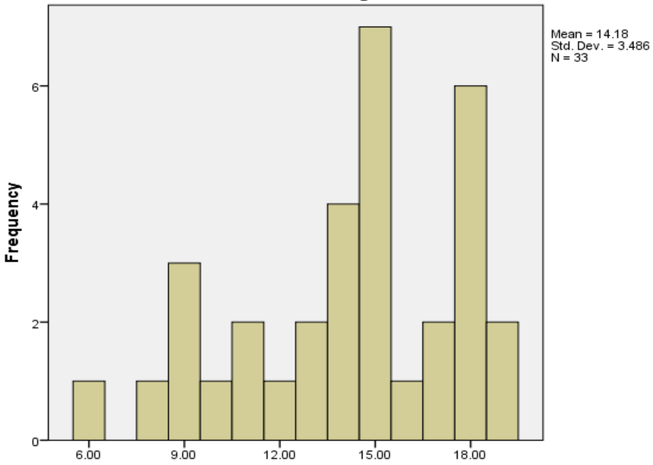
**TABLE 2**
**Descriptive Statistics for the MC Test of Pragmatic Knowledge**

| | *Valid N* | *Range* | *Mean* | *Median* | *SD* | *Skewness/SD* | *Kurtosis/SD* |
|---|---|---|---|---|---|---|---|
| Measure | 33 | 13 | 14.18 | 15 | 3.48 | -.58 / .40 | -.46 / .79 |

### *Measures of Central Tendency*

The mean for the test was 14.18 (out of a possible total score of 21 points) and a median of 15. These estimates are slightly higher than would be desired for a proficiency test, but it does not seem that the test was overly easy; on average, test-takers responded correctly to 67.52% of the items. The standard deviation of 3.48 indicates that there was variability among the test-takers. The skewness was -.58, and kurtosis was -.46. Given the proximity of the mean and median, and the fact that both skewness and kurtosis estimates fell within the desired range, it can be argued that the distribution resembles normality. The negative skewness index indicates that the test might have been easy for the sample that took it, in particular when considering that the test was designed as a proficiency test of pragmatic knowledge and it aimed to distinguish between different levels of ability. Figure 3 shows the distribution of the test.

**FIGURE 3**
**Distribution of overall scores for the test**



## Reliability Analyses

The test had a Cronbach's Alpha coefficient of .68. This indicates that 68% of score variance can be attributed to true score variance, leaving 32% of the score as due to measurement error or construct-irrelevant variance. While this indicator falls short from acceptable standards for internal consistency, this is a solid start for the piloting stage, in particular considering that several studies on pragmatics have had very low internal-consistency measures (e.g., Roever, 2010), specifically where selected-response tasks were concerned. It should be noted that one item was excluded from the analysis since it presented no variability (Item 15).

## Dichotomous Rasch Model Analysis

In order to determine whether the test successfully separated lower-and high-ability learners, Rasch analysis was conducted to analyze the distribution and separation of items (in terms of difficulty) and persons (in terms of ability). The assumption of unidimensionality was examined as a precondition for Rasch Analysis. The Principal Components Rasch Analysis (PCRA) plot showed no signs of heavy clustering (see Appendix D), indicating the satisfaction of the precondition. These analyses, however, are only exploratory and should be taken with caution, since the sample size is too small to make any definitive claims about the results.

### *Coverage and Item Separation*

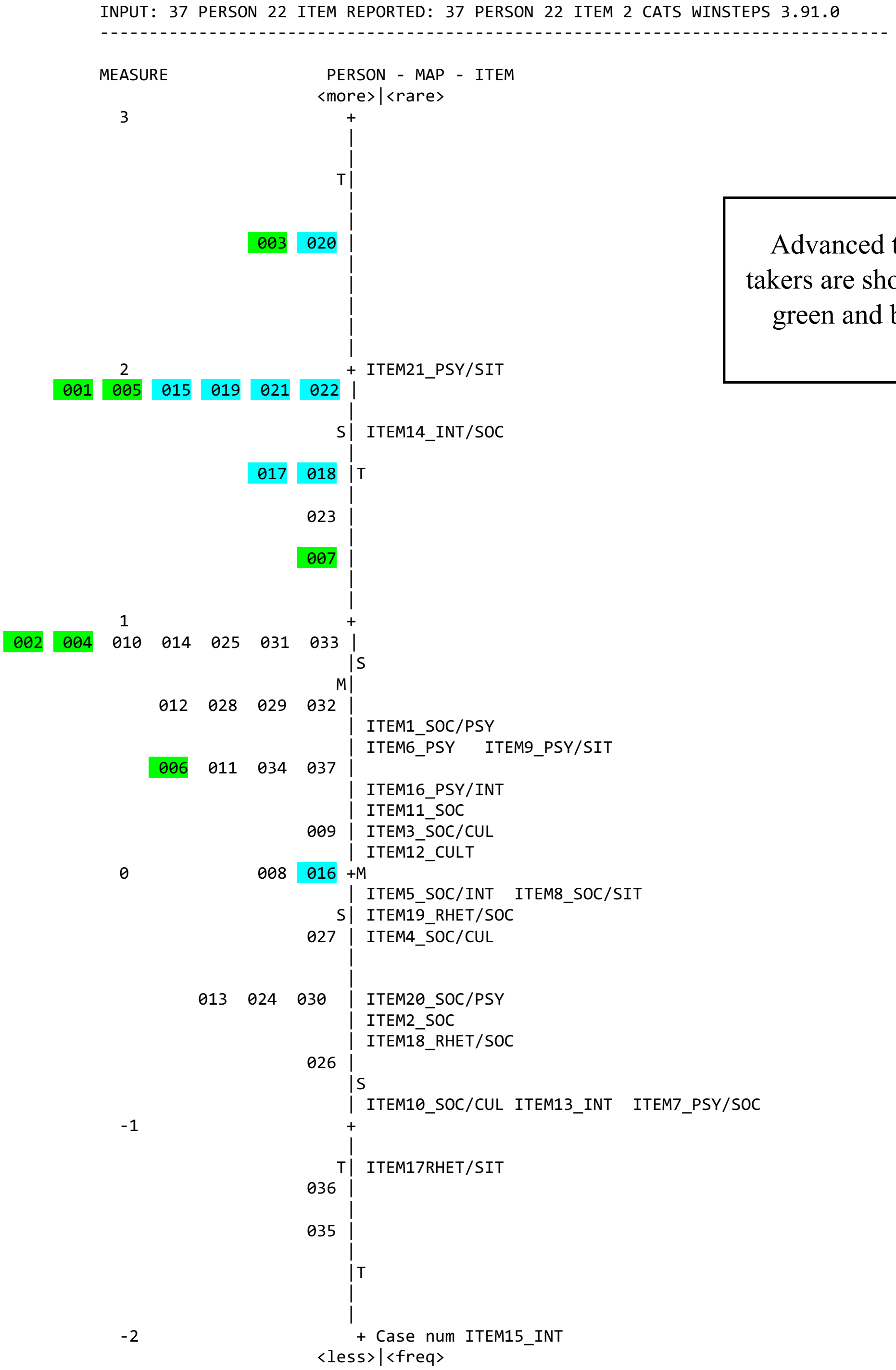
As a reminder, in Rasch Analysis items and persons are placed on an interval scale, that is, a logit scale. Regarding item difficulty, most items seem to be located within 2 and -1 logits, which shows a logit spread of 3 logits. The only exception to this range is Item 17, which was located at -1.17 logits. These numbers become more meaningful upon examination of the Wright map produced by *WINSTEPS.* As can be seen, the map places examinees and items on two histograms on the logit scale, which in this case is shown to range from -2 to 3. Given that the Rasch model examines persons and items, there are only two sets of information. On the left side, test-takers (which are displayed through ID numbers from 001 to 037) are placed on the scale according to their estimated pragmatic ability. The higher-ability test-takers are on top, and the lower ability test-takers are displayed at the bottom. In the case of items, which are displayed on the right, those that are placed at the bottom of the map are easier items, while those on the top are the most difficult ones. The map shows a mismatch between item difficulty and examinee ability, since examinees appear to have a wider spread (ranging between -1.5 to 2.3) than the items (ranging between -1 to 2). This indicates that there is no adequate coverage for the higher ability test-takers in this test of pragmatic knowledge (those positioned above 2 in the logit scale), implying that the test has limited power to discriminate among the higher-ability test takers. This issue should be revisited in order to overcome the issues of ceiling effects, whether by increasing the pool of items with more complex ones or revising existing items to increase their difficulty.

The Item Separation Index was found to be very low (G=1.84), which would indicate that the test does not possess a wide range of item difficulty. This implies that items with different qualities should be included in future iterations of the test. However, given that the reliability of the separation was only .77, it might also be possible that the sample size is not large enough to confirm the item hierarchy (Linacre, 2019). Strata were calculated at 7.36 (Strata= 4*1.84 + 1 = 7.36), which shows that there are about seven statistically distinct levels of item difficulty that separated test-takers by ability by at least three errors of measurement (Wright & Masters, 2002). Although item separation was found to be lower than acceptable, it should be noted that a larger sample would be required to provide a more solid argument. As of now, it would seem that the test does not adequately target different levels of ability, since most items are located within a narrow range of difficulty.

### *Item Fit*

In order to flag possibly problematic items, the Standardized Infit and Outfit Mean Square columns were examined since the sample size is smaller than 200. According to Bond and Fox's (2015) recommended range for model fit, all items fall within the -2 and 2 values for

## FIGURE 4
## WINSTEPS summary: Wright Plot, including the calibration of test items



```
        INPUT: 37 PERSON 22 ITEM REPORTED: 37 PERSON 22 ITEM 2 CATS WINSTEPS 3.91.0
        --------------------------------------------------------------------------------

        MEASURE                PERSON - MAP - ITEM
                              <more>|<rare>
           3                      +
                                  |
                                  |
                                 T|
                                  |
                                  |
                         003  020 |
                                  |
                                  |
                                  |
                                  |
                                  |
           2                      + ITEM21_PSY/SIT
     001  005  015  019  021  022 |
                                  |
                                 S| ITEM14_INT/SOC
                                  |
                         017  018 |T
                                  |
                              023 |
                                  |
                              007 |
                                  |
                                  |
           1                      +
002  004  010  014  025  031  033 |
                                  |S
                                 M|
               012  028  029  032 |
                                  | ITEM1_SOC/PSY
                                  | ITEM6_PSY   ITEM9_PSY/SIT
               006  011  034  037 |
                                  | ITEM16_PSY/INT
                                  | ITEM11_SOC
                              009 | ITEM3_SOC/CUL
                                  | ITEM12_CULT
           0             008  016 +M
                                  | ITEM5_SOC/INT  ITEM8_SOC/SIT
                                 S| ITEM19_RHET/SOC
                              027 | ITEM4_SOC/CUL
                                  |
                                  |
                   013  024  030  | ITEM20_SOC/PSY
                                  | ITEM2_SOC
                                  | ITEM18_RHET/SOC
                              026 |
                                  |S
                                  | ITEM10_SOC/CUL ITEM13_INT  ITEM7_PSY/SOC
           -1                     +
                                  |
                                 T| ITEM17RHET/SIT
                              036 |
                                  |
                              035 |
                                  |
                                  |T
                                  |
                                  |
           -2                      + Case num ITEM15_INT
                              <less>|<freq>
```

Advanced test-takers are shown in green and blue

the Standardized Infit. Most items showed little erratic response behavior for non-extreme observations (Linacre, 2019), with only two items at borderline levels: Items 5 at 1.34 and 19 at

.67, which fell slightly outside recommended parameters of .7-1.3 Infit (Bond & Fox, 2015). However, when examining the Outfit MNSQ indices, one item falls at 2.0 (Item 2), showing a possible tendency to be underfitting due to erratic behavior (Bond & Fox, 2015). Therefore, this item should be more closely examined and either be revised or dropped from the test, especially considering that the classical test theory analysis indicated that removing this item would increase Cronbach's Alpha to .72. To view the complete set of item infit and outfit indices, see Appendix E.

## Multiple Regression Analysis

In order to determine whether the background variables of length of stay, educational background, proficiency, and L1 had an impact on test-takers' PKC, linear multiple regression analysis was employed. Multiple regression analysis is used to examine the linear relationship between two or more variables, where one variable (*Y*) is believed to be dependent on other (independent) variables (*X*'s). This analysis attempts to calculate *Y* from a given number (*k*) of regressors, which produces partial regression weights (b). Each partial regression weight indicates how much change that variable (*X*) produces on Y when everything else is held constant (Darlington & Hayes, 2016). In this study, there were four predictors (also known as regressors) for the regression model for Total Score of PKC (*Y*): First Language($X_1$), Length of Stay($X_2$), Time Studying English($X_3$), and Educational Degree ($X_4$), so that $k=4$. Each of these predictors is assigned a partial regression weight (ß) indicating how much change they exert on *Y*. The regression equation for this study is the following:

$$\underset{\text{Total Score PKC}}{Y=} \quad \underset{\text{Intercept}}{\beta_0} \quad \underset{\text{First Language}}{\beta_1 X_1 +} \quad \underset{\text{Length of Stay}}{\beta_2 X_2 +} \quad \underset{\text{Time Studying English}}{\beta_3 X_3 +} \quad \underset{\text{Educational Degree}}{\beta_4 X_4 +} \quad \underset{\text{Error}}{\mathcal{E}}$$

In addition, composite scores for each of the subcomponents of pragmatic knowledge were calculated in order to use multiple regression analysis with each subcomponent (for each of these analyses, the subcomponent score was the dependent variable *Y*). For example, for Sociocultural Knowledge, the composite score could have a maximum score of 4 (after adding up items 3, 4, 10, and 12, all of which targeted this subcomponent). Table 4 shows the results for the first model under scrutiny, which included all four background variables as predictors of PKC.

The proposed model for predicting pragmatic knowledge control was comprised of four regressors: Native Language (L1), Length of Stay (LOS), Time Studying English, and Last Completed Educational Degree (EdDegree). The model reached an $R^2$ of .332, which indicates that 33.2% of the variance in PKC can be explained by the model (F=3.35, *p*.=.02). However, it was found that only one of the variables held a significant relationship (*p*.<.001) with the dependent variable of pragmatic knowledge control: Time Studying English. Regarding this variable, it was found that, as test-takers' amount of years studying English increases by one unit, their score on the test of pragmatic knowledge increases by .28 points (in the test scale).

**TABLE 4**
**Coefficients[a] of the multiple regression analysis**

| | Model | *Unstandardized Coefficients* | | *Standardized Coefficients* | | |
|---|---|---|---|---|---|---|
| | | *B* | *Std. Error* | *Beta* | *t* | *Sig.* |
| 1 | (Constant) | 11.281 | 1.953 | | 5.775 | .000 |
| | L1 | .263 | .279 | .152 | .943 | .354 |
| | LOS | -.001 | .013 | -.014 | -.085 | .933 |
| | TimeStudy | .285 | .079 | .590 | 3.620 | .001 |
| | EdDegree | -.356 | .707 | -.081 | -.504 | .619 |

a. Dependent Variable: Total_Test_Score

In order to determine if there were any differences in how these predictors functioned in relation to test-takers' proficiency level, the model was run for each of the subgroups of the sample according to proficiency. In this way, proficiency was operationalized as the program labels of intermediate and advanced students, so that test-takers were coded as advanced (1) and intermediate (2) in the factor/categorical variable of proficiency. Tables 5 and 6 show the results of the model. As can be seen in the tables, when accounting for proficiency groupings, Last Educational Degree became a significant predictor for both advanced and intermediate students.

**TABLE 5**
**Coefficients[a,b] of the multiple regression analysis for advanced examinees**

| | *Model* | *Unstandardized Coefficients* | | *Standardized Coefficients* | | |
|---|---|---|---|---|---|---|
| | | *B* | *Std. Error* | *Beta* | *t* | *Sig.* |
| 1 | (Constant) | 16.598 | 1.848 | | 8.981 | .000 |
| | L1 | -.075 | .227 | -.077 | -.330 | .749 |
| | LOS | -.021 | .027 | -.180 | -.763 | .463 |
| | TimeStudy | .291 | .069 | .964 | 4.232 | .002 |
| | EdDegree | -1.712 | .643 | -.597 | -2.661 | .024 |

a. Dependent Variable: Total_Test_Score
b. Selecting only cases for which Proficiency = Advanced

**TABLE 6**
**Coefficients[a,b] of the multiple regression analysis for intermediate examinees**

| | *Model* | *Unstandardized Coefficients* | | *Standardized Coefficients* | | |
|---|---|---|---|---|---|---|
| | | *B* | *Std. Error* | *Beta* | *t* | *Sig.* |
| 1 | (Constant) | 8.973 | 2.346 | | 3.825 | .002 |
| | L1 | -.723 | .531 | -.411 | -1.362 | .198 |
| | LOS | -.007 | .013 | -.133 | -.541 | .599 |
| | TimeStudy | -.123 | .164 | -.197 | -.751 | .467 |
| | EdDegree | 2.760 | 1.168 | .732 | 2.363 | .036 |

a. Dependent Variable: Total_Test_Score

b. Selecting only cases for which Proficiency= Intermediate

As shown in Tables 5 and 6, when proficiency is accounted for, there seems to be a difference in how the background variables interact with the total score in the regression model. For the advanced learners, in addition to Time Studying English, the variable of Last Educational Degree was also found to be a significant predictor of pragmatic ability. While Time Studying English maintained a similar coefficient to the model without groupings, the variable of Last Educational Degree had a beta coefficient of -1.71, which indicates an inverse relationship with pragmatic knowledge control. This indicates that for each one-unit increase in the category of Last Educational Degree, there is a decrease of 1.71 points on test-takers' scores when everything else is held constant.

In contrast, the intermediate group had only one statistically significant predictor of pragmatic knowledge control, Last Educational Degree, which had a direct relationship with pragmatic knowledge control. For each category increase in Last Educational Degree, intermediate test-takers score 2.76 units higher in the test when everything else is held constant. In other words, there is a difference of 2.76 units in the total score of intermediate test-takers from different groupings for Last Educational Degree. Those intermediate test-takers who earned a bachelor's degree are predicted to have a score that is 2.76 higher than their counterparts who only have a high school diploma. Last Educational Degree is the only significant predictor for intermediate test takers, and as determined by $R^2$, the model explained only 27.3% of the variance in the total test scores.

These results indicate that, for the current sample, test-takers' Last Educational Degree is the best predictor of their overall score when proficiency is accounted for in the regression model. This suggests that formal education might be an important predictor of PKC (given the two significant predictors are related to language study and overall academic experience, which possibly have a role to play in this type of instrument). Furthermore, the results indicate that Last Educational Degree contributes differently to the total score on PKC depending on general language proficiency. Nonetheless, it should be noted that there could be other variables that would make better predictors of PKC scores but that were not explored in this study.

In order to determine whether each subcomponent of PKC from Purpura's (2017) adapted model had any interaction with the background variables, multiple regression was performed for each of the subscores of the test (i.e., for each of the composite scores comprised of all items targeting a given type of pragmatic meaning). The goal was to determine whether there were any differences from the overall model of PKC. However, the only statistically significant predictor was, once again, Time Studying English. Hence, Time Studying English was a statistically significant predictor for the subcomponents of: Situational, Sociolinguistic, Psychological, and Rhetorical Knowledge. To see the specific coefficients for each of these subcomponents, refer to Appendix F.

When taking into consideration the proficiency groupings, it was found that Rhetorical Knowledge had an additional significant predictor for PKC for advanced test-takers: the variable of Last Educational Degree. Given that the Educational Degree variable was categorical, it should be understood that there is a decrease of .553 in the predicted rhetorical knowledge subscores for each one-unit difference in Last Educational Degree when everything else is held constant (note that the categories were in ascending order). This means, for example, that advanced learners with a doctoral degree had a predicted rhetorical knowledge subscore that was .553 points lower than the score of advanced test-takers with only a master's degree. Interestingly, as determined by $R^2$, the model explained 77.1% of the variance in the rhetorical knowledge sub-scores, which is a large amount of variation explained by the regressors. A simplified model which included only Time Studying English and Last Educational Degree had an $R^2$=.64, showing that the combination of these two variables accounts for a large amount of score variability even when the other two regressors are deleted from the model for advanced test-takers. Next, the implications of these findings are discussed.

**TABLE 7**
**Regression Coefficients[a,b] for Rhetorical Knowledge (Advanced)**

| | | *Unstandardized Coefficients* | | *Standardized Coefficients* | | |
|---|---|---|---|---|---|---|
| *Model* | | *B* | *Std. Error* | *Beta* | *t* | *Sig.* |
| 1 | (Constant) | 3.347 | .382 | | 8.758 | .000 |
| | L1 | -.080 | .047 | -.333 | -1.695 | .121 |
| | LOS | -.013 | .006 | -.448 | -2.256 | .048 |
| | TimeStudy | .076 | .014 | 1.028 | 5.344 | .000 |
| | EdDegree | -.553 | .133 | -.788 | -4.155 | .002 |

a. Dependent Variable: Rhetorical
b. Selecting only cases for which Proficiency = Advanced

## DISCUSSION

In order to discuss the implications of these findings, each of the research questions will be addressed in turn.

1. *How can a meaning-based model of pragmatic ability be used as the basis for the design of a pragmatic ability test?*

Throughout the development of this test, Purpura's (2017) model was used as the theoretical model underlying the construct of pragmatic knowledge. One advantage of using this model was that it was possible to go beyond the typical characterizations of pragmatic knowledge in terms of imposition, social distance, and power differential, or from the dichotomy of sociopragmatics and pragmalinguistics (cf. Leech, 1983), opting for contextually derived meanings that broaden the conceptualization of pragmatic knowledge. Moreover, a clear advantage is that this model allows for the differentiation of implied pragmatic meanings from purely grammatical/lexical meanings (as distinguished in Purpura, 2017). However, the limited range of item difficulties shows one caveat that often surfaces in the exploration of new models, and that is, that the theoretical difficulty of an item will not always match those that are found empirically, just as it cannot be ensured that the items are eliciting relevant knowledge, skills, and abilities from test-takers (Roever & McNamara, 2006). Thus, in addition to sampling a wider range of situations to elicit displays of understanding of different types of knowledge, it is necessary to analyze the features that may affect item difficulty, which ties into the second research question of the study.

2. *How effectively can a MC test of pragmatic knowledge separate test-takers in terms of their ability?*

Results from the dichotomous Rasch analyses show that the test had a limited power to discriminate between high and low-ability test takers. The main problem, it was found, was a ceiling effect, which showed a mismatch between the distribution of items and that of test-takers by ability level. That is, the test did not include enough items that targeted the high-ability test-takers. Similarly, some items seemed to be too easy. For instance, one of the items was excluded from the analysis due to its lack of variability, which is another indicator of the need to revisit item difficulty. On the other hand, most of the distribution of items regarding the subcomponents of pragmatic knowledge showed that it is possible to have items of varying difficulties for most subcomponents. The only subtype of knowledge that remained below the *0* logit (and thus within a very narrow spread of item difficulty) was rhetorical knowledge. However, this is partially explained by the fact that there were fewer items for this subcomponent (only three). Therefore, revising the test is recommended, in particular to increase its discriminatory power. Such revisions would include revisiting the distractors for those items that had high facility, as well as revisiting the keys to avoid excessive cueing or prompting.

3. *To what extent is there a relationship between test-takers' time studying English, academic background, length of stay in the US, and L1 background with pragmatic knowledge?*

Mixed results were found in relation to this question. No meaningful relationship was found between length of stay in the US or L1 background with test takers' pragmatic knowledge. This finding is in line with Kasper and Rose's (2002) depiction of length of stay as an unreliable predictor of pragmatic ability. The only predictor to consistently predict pragmatic knowledge was the number of years studying English, which was a statistically significant predictor both for the overall test score and for the subscores for Situational, Sociocultural, Psychological, and Rhetorical Knowledge.

However, it was also found that when taking into account proficiency groupings, test-takers' academic background had an effect on predicting pragmatic knowledge. Increased educational degrees favored intermediate test-takers, while the inverse relationship was found for advanced test-takers in their pragmatic knowledge scores. Rhetorical knowledge scores were also found to be (inversely) predicted by Educational Degree when proficiency was accounted for, for the advanced group (but this time the regressor was not significant for intermediate test-takers). There might be various reasons for this inverse relationship. It could be argued that students with higher-education degrees who are still enrolled in ESL classes do so because of their ongoing struggles learning the language. In the case of pragmatic knowledge control, it could also be that students with advanced degrees tend to be overly formal due to the lack of exposure to more informal settings in the target culture. However, these are speculations, and the negative relationship for the advanced test-takers could be sample-specific and circumstantial. For example, in the advanced group, there were only two test-takers with only a high school degree and two with a master's degree. Yet, only one of the students with a master's degree had a score of one out of three for rhetorical meanings. This might have resulted in the observed partial regression weights, so that having a more diverse sample would probably yield different results. Thus, given the problem of sample size, these results have very limited generalizability, and the relationships are to be considered with caution. This would need to be studied further with a larger, more diverse sample to overcome the constraints of interpretation of the effect of Educational Degree on PKC as suggested by the current study.

## CONCLUSION

In conclusion, while a meaning-based model (Purpura, 2017) seems to be a viable option to further study the context of pragmatic ability, numerous actions need to be taken in terms of revisiting item development and overall test design. Including only one type of task might have had a negative effect on the measurement qualities of the test, particularly when it comes to item separation and the test's power to discriminate between high-and low ability test-takers. It is hoped that revisiting item difficulty, both from a theoretical and empirically derived standpoint, would provide useful information for the construct validation of pragmatic knowledge as proposed in this study.

One major limitation of this study is that the sample size was small, and, therefore, the findings are to be viewed with caution, in particular given the small statistical power of these analyses. In addition, the results could only be generalizable to similar contexts and populations. Moreover, different sampling arrangements should be made to have similar proportions of items for each of the subcategories under comparison. Accordingly, a replication of this study (following thorough revision of the test, particularly revising the items that had little variance) would highly benefit from increasing the sample of test takers. Furthermore, it would allow for the use of more powerful statistical analyses, such as DIF, once the minimum sampling is satisfied for the dichotomous Rasch model. In addition, implementing Structural Equation Modeling would allow for the empirical examination of the construct validity of the meaning-based model of pragmatic knowledge.

## REFERENCES

Bardovi-Halig, K., & Bastos, M. T. (2011). Proficiency, length of stay, and intensity of interaction, and the acquisition of conventional expressions in L2 pragmatics. *Intercultural Pragmatics 8*(3), 347–384.

Barron, A. (2003). *Second language acquisition in a study abroad context*. Amsterdam, Netherlands: Benjamins.

Bataller, R. (2010). Making a request for a service in Spanish: pragmatic development in the study abroad setting. *Foreign Language Annals, 43*(1), 160–175.

Blum-Kulka, S., & Olshtain, E. (1986). Too many words: Length of utterance and pragmatic failure. *Studies in Second Language Acquisition, 8*(2), 165–180.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New York, NY, US: Routledge.

Cohen, A. D., & Shively, R. L. (2007). Acquisition of requests and apologies in Spanish and French: Impact of study abroad and strategy-building intervention. *The Modern Language Journal, 91*, 189–212.

Darlington, R. B. & Hayes, A. F. (2016). *Regression analysis and linear models: Concepts, applications, and implementation*. New York, USA: The Guilford Press.

Downing, S. M. (2003). Item response theory: Applications of modern test theory for assessments in medical education. *Medical Education, 37*, 739–745.

Félix-Brasdefer, J. C. (2004). Interlanguage refusals: Linguistic politeness and length of residence in the target community. *Language Learning, 54*(4), 587–653.

Félix-Brasdefer, J. C. (2007). Pragmatic development in the Spanish as a FL classroom: A cross-sectional study of learner requests. *Intercultural Pragmatics, 4*, 253–286.

Grabowski, K. (2009). *Investigating the construct validity of a test designed to measure grammatical and pragmatic knowledge in the context of speaking.* (Unpublished doctoral dissertation). Teachers College, Columbia University, New York, NY.

Granger, C. V. (2008). Rasch analysis is important to understand and use for measurement. *Rasch Measurement Transactions, 21*(3), 1122–1123.

Hoffman-Hicks, S. (1999). *The longitudinal development of French foreign language pragmatic competence: Evidence from study-abroad participants* (Unpublished doctoral dissertation). Indiana University, Bloomington.

IBM Corp. (2017). *IBM SPSS Statistics for Windows, Version 25.0*. Armonk, NY: IBM Corp.

Kasper, G., & Rose, K. R. (2002). *Pragmatic development in a second language*. Malden: Blackwell.

Kobayashi, H., & Rinnert., C. (2003). Coping with high imposition requests: High vs. low proficiency EFL students in Japan. In A. Martínez, E. Usó Juan, & A. Fernández (Eds.), *Pragmatic competence and foreign language teaching* (pp. 161–184). Castellon, Spain: Servei de Publicacions de la Univerisitat Jaume I.

Leech, G. N. (1983). *Principles of pragmatics.* London, England: Longman.

Linacre J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(2), 328. Retrieved from: https://www.rasch.org/rmt/rmt74m.htm

Linacre, J. M. (2015). *Winsteps®* (Version 3.91.0) [Computer Software]. Beaverton, Oregon: Winsteps.com. Retrieved January 30, 2019. Available from http://www.winsteps.com

Linacre, J. M. (2019, May) Reliability and separation of measures. Retrieved from https://www.winsteps.com/winman/reliability.htm

Mislevy, R., Almond, R., & Lukas, J. (2004). *A brief introduction to Evidence-Centered Design, CSE Report 632.* Los Angeles: Center for Research on Evaluation, Standards, & Student Testing, US Department of Education.

Nguyen, T. M. 2008. Modifying L2 criticisms: How learners do it? *Journal of Pragmatics 40*, 768–791.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge, UK: Cambridge University Press.

Purpura, J. E. (2017). Assessing meaning. In E. Shohamy, L. Or, & S. May (Eds.), *Language testing and assessment: Encyclopedia of language and education* (pp. 33–61). New York, NY: Springer International Publishing. doi: 10.1007/978-3-319-02326-7_1-1

Roever, C. (2007). DIF in the assessment of second language pragmatics. *Language Assessment Quarterly, 4*(1), 165–187.

Roever, C. (2010). Effects of cultural background in a test of ESL pragmalinguistics: A DIF approach. In G. Kasper, H. T. Nguyen, D. R. Yoshimi, & J. K. Yoshioka (Eds.), *Pragmatics language learning* (Vol. 12, pp. 187–212). Honolulu, HI: National Foreign Language Resource Center, University of Hawai'i at Manoa.

Roever, C. (2011). Testing of second language pragmatics: Past and future. *Language Testing 28*(4), 463–481.

Roever, C. (2014). Assessing pragmatics. In A. J. Kunnan (Ed.), *The companion to language assessment* (1st ed., pp. 125-139). Somerset, NJ: John Wiley & Sons, Inc. doi: 10.1002/9781118411360.wbcla057

Roever, C., & McNamara, T. (2006). Language testing: The social dimension. *International Journal of Applied Linguistics, 16*(2), 242–258.

Schauer, G. A. (2007). Finding the right words in the study abroad context: The development of German learners' use of external modifiers in English. *Intercultural Pragmatics 4*, 193–220.

Shively, R. L. (2011). L2 pragmatic development in study abroad: A longitudinal study of Spanish service encounters. *Journal of Pragmatics, 43*(6), 1818–1835.

Timpe-Laughlin, V., Wain, J., & Schmidgall, J. (2015). *Defining and operationalizing the construct of pragmatic competence: Review and recommendations*. (Research Memorandum No. *RR-15-06*). Princeton, NJ: Educational Testing Service.

VonCannon, A. L. (2006). *Just saying 'no': Refusing requests in Spanish as a first and second language* (Unpublished doctoral dissertation). University of Iowa, Iowa City, IA.

Wiersma, W., & Jurs, S. G. (2009). *Research Methods in Education. An Introduction*. Boston, MA: Allyn & Bacon.

Wright, B. D., & Masters, G. N. (2002). Number of person or item strata: (4*Separation+1)/3. *Rasch Measurement Transactions, 7*(2), 328. Retrieved from: https://www.rasch.org/rmt/rmt163f.htm

## APPENDIX A

## Sample Items from the Pragmatics Test

In each situation, select the best answer.

1. **Item 1. You are in an Economics class. Towards the end of the class, you ask your professor to explain again one of the topics that you covered earlier. As your professor finishes his explanation, you still don't understand the key concept, but have decided you will continue to study at home.**

Professor: ... And, therefore, that is the reason why you would have to consider both theories. Is that clear?
You: _______________________.

a. Mmm, I'll just study at home.
b. I still don't get it, but thanks anyway. I'll study at home.
c. Yeah, I guess. It's not your explanation, though. Maybe I should just go over this at home.
d. I'm still processing it, but I don't want to take any more of your time, so I'll go over this at home.

2. **Item 4. You are in the second week of a creative writing course. You are in class. All students are working on a short writing assignment. Your professor, Dr. Thomas**

**Spence, insists that all students call him Tom. You have been trying to do so, but you sometimes forget about this preference**.

You: Professor, can I ask a question?
Him: Yes, student.
You: _________________________.

a. Oh, sure, sure. Dr. Tom. I totally forgot.
b. Oh, right. I'm still getting used to it, Tom.
c. Well, I don't think professors should go by first name.
d. I forgot, but I'd rather not call you Tom, it's just not my style.

3. **Item 5. You are waiting in line to pay for a soda at the school cafeteria. When it is your turn, the cashier starts chatting with the barista, and they don't seem to be finishing any time soon.**

How would you get their attention?

a. Excuse me, can I pay for this?
b. Would it be OK if I paid for this?
c. Erm...I don't mean to interrupt, but can I pay for this?
d. Hey, I've been waiting for a while. Can I pay for this?

**8-9. Items 8-9. Look at how the conversation develops. What would you say?**

Clerk: Ok, let me check your records. What's your ID number?
You: 556239.
Clerk: Ok...one second...mmm, it seems that you are billing your tuition to a third party, do you have a scholarship from another institution?
You: Yes, and my tuition is supposed to be paid for, I received an invoice last month.
Clerk: I see. Unfortunately, the person who takes care of scholarship payments is out today. She'll be in tomorrow, though.

You: ___________________________________________________.

a. Well, I'm not coming to school tomorrow.
b. Can you assure me she'd be here tomorrow?
c. Why wouldn't she be here today? It's her job.
d. Can I speak to her manager? I'm not coming again.

Clerk: Well, I'm sorry that we can't help you now, but, you don't even have to come in person. You can just give us a call and say that you need to talk to Claire, the scholarships manager.

You: ___________________________________________________.
a) The phone it is then. But I'm definitely not coming back in person.
b) I guess I'll make the call then. I really wanted to get this sorted out today, though.

c) I’m going to call the office tomorrow, but I am not leaving until the manager sees me.
d) So, I’ll call the office then. It’s a shame she isn’t here today, though. It’s my last day coming to school this week.

Clerk: I’m sure this can be sorted out promptly without another visit. Is there anything else you need?

# APPENDIX B

## Sample Answer Key and Item Coding

### Answer Key and Coding

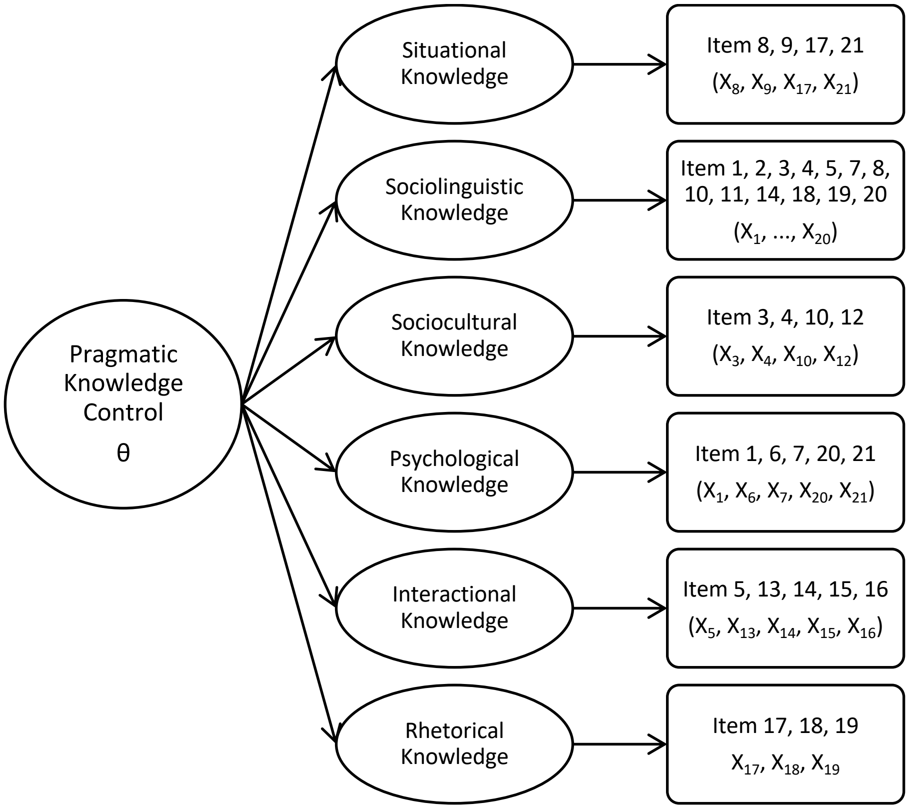
| | | |
|---|---|---|
| 1 | D | This option conveys the most appropriate sociolinguistic and psychological meanings. (SOC + PSY) |
| 4 | B | Option B has the most adequate response given the previous request of the professor to go by first name and the sociocultural norms in terms of classroom culture. (SOC, CUL) |
| 7 | C | This option has the best tone when addressing the clerk. (PSY, SOC) |
| 8 | A | Option A makes use of implicature to indirectly communicate the inconvenient. (SOC, SIT) |
| 9 | B | Option B best conveys lack of satisfaction without an overly negative/impolite response. (PSY, SIT) |

### Answer Key

| | |
|---|---|
| 1 | D |
| 4 | B |
| 7 | C |
| 8 | A |
| 9 | B |

# APPENDIX C

## Operational Model of Pragmatic Knowledge Control


Pragmatic Knowledge Control
θ
Situational Knowledge
Item 8, 9, 17, 21
($X_8$, $X_9$, $X_{17}$, $X_{21}$)
Sociolinguistic Knowledge
Item 1, 2, 3, 4, 5, 7, 8, 10, 11, 14, 18, 19, 20
($X_1$, …, $X_{20}$)
Sociocultural Knowledge
Item 3, 4, 10, 12
($X_3$, $X_4$, $X_{10}$, $X_{12}$)
Psychological Knowledge
Item 1, 6, 7, 20, 21
($X_1$, $X_6$, $X_7$, $X_{20}$, $X_{21}$)
Interactional Knowledge
Item 5, 13, 14, 15, 16
($X_5$, $X_{13}$, $X_{14}$, $X_{15}$, $X_{16}$)
Rhetorical Knowledge
Item 17, 18, 19
$X_{17}$, $X_{18}$, $X_{19}$

# APPENDIX D

## Variance Component Scree Plot

```
   +--+--+--+--+--+--+--+--+--+--+--+
 100%+ T               +
  |                 |
V 63%+       U         +
A  |                |
R 40%+                +
I  |                |
A 25%+                +
N  |   M             |
C 16%+                +
E  |      I          |
 10%+    P    1       +
L  |          2      |
O 6%+            3    +
G  |             4 5 |
| 4%+                +
S  |                |
C 3%+                +
A  |                |
L 2%+                +
E  |                |
D 1%+                +
  |                 |
 0.5%+                +
   +--+--+--+--+--+--+--+--+--+--+--+
    TV MV PV IV UV U1 U2 U3 U4 U5
    VARIANCE COMPONENTS
```

# APPENDIX E

## Item Statistics: Misfit Order (Outfit and Infit)

```
-----------------------------------------------------------------------------------------------
|ENTRY  TOTAL TOTAL      MODEL|  INFIT | OUTFIT |PTMEASUR-AL|EXACT MATCH|      |
|NUMBER SCORE COUNT MEASURE S.E. |MNSQ ZSTD|MNSQ ZSTD|CORR. EXP.| OBS% EXP%| ITEM    |
|------------------------------------+----------+----------+-----------+-----------+-----------|
|   2   28   37  -.58   .42|1.26  1.2|1.81  2.0|A .06  .38| 73.0 77.8| ITEM2   |
|   5   25   37  -.10   .38|1.34  1.9|1.37  1.4|B .09  .40| 59.5 73.0| ITEM5   |
|  14   11   34  1.74   .40|1.26  1.5|1.16  .7|C .15  .37| 55.9 71.5| ITEM14   |
|   9   20   36   .52   .37|1.22  1.5|1.23  1.2|D .22  .42| 58.3 68.4| ITEM9   |
|   3   23   37   .18   .37|1.18  1.2|1.19  .9|E .25  .41| 64.9 70.5| ITEM3   |
|   6   20   36   .52   .37|1.13  .9|1.13  .7|F .31  .42| 63.9 68.4| ITEM6   |
|  13   29   36  -.91   .45|1.13  .6| .95  .0|G .28  .36| 80.6 81.6| ITEM13   |
|  20   25   33  -.51   .44|1.03  .2| .85  -.3|H .38  .37| 75.8 77.1| ITEM20   |
|  16   20   33   .33   .39|1.02  .2|1.02  .2|I .39  .41| 69.7 69.7| ITEM16   |
|  21   9   33  2.00   .42| .99  .0| .95  .0|J .37  .36| 66.7 74.0| ITEM21   |
|  11   22   36   .25   .38| .95  -.3| .86  -.6|j .48  .42| 66.7 70.4| ITEM11   |
|  10   29   36  -.91   .45| .88  -.4| .91  -.1|i .44  .36| 86.1 81.6| ITEM10   |
|   4   26   37  -.25   .39| .88  -.6| .75  -.9|h .53  .40| 78.4 74.5| ITEM4   |
|   8   24   36  -.04   .39| .88  -.7| .77  -.9|g .54  .41| 75.0 72.8| ITEM8   |
|  18   26   33  -.71   .46| .88  -.4| .70  -.7|f .49  .35| 78.8 79.6| ITEM18   |
|   7   29   36  -.91   .45| .85  -.5| .58 -1.0|e .54  .36| 80.6 81.6| ITEM7   |
|  12   23   36   .10   .38| .85  -.9| .79  -.9|d .56  .42| 77.8 71.6| ITEM12   |
|   1   20   37   .59   .36| .83 -1.3| .79 -1.2|c .58  .42| 81.1 67.8| ITEM1   |
|  17   28   33  -1.17   .51| .77  -.7| .53  -.9|b .56  .31| 87.9 84.9| ITEM17   |
|  19   23   33  -.15   .41| .68 -1.8| .58 -1.7|a .70  .39| 87.9 73.7| ITEM19   |
|------------------------------------+----------+----------+-----------+-----------+-----------|
| MEAN  22.5  33.6  -.15   .56|1.00  .1| .95  -.1|      | 73.4 74.5|      |
| P.SD   7.3  7.3  1.21   .47| .18  1.0| .30  1.0|      | 9.5  5.0|      |
-----------------------------------------------------------------------------------------------
```

# APPENDIX F

## Multiple Regression Results for PKC Subcomponents

### Situational Knowledge Regression Coefficients[a]

| | | *Unstandardized Coefficients* | | *Standardized Coefficients* | | |
|---|---|---|---|---|---|---|
| *Model* | | B | Std. Error | Beta | *t* | *Sig.* |
| 1 | (Constant) | 1.056 | .574 | | 1.841 | .077 |
| | L1 | .151 | .082 | .296 | 1.839 | .077 |
| | LOS | .002 | .004 | .075 | .474 | .639 |
| | TimeStudy | .078 | .023 | .546 | 3.368 | .002 |
| | EdDegree | .001 | .208 | .001 | .005 | .996 |

a. Dependent Variable: Situational

### Sociolinguistic Knowledge Regression Coefficients[a]

| | | *Unstandardized Coefficients* | | *Standardized Coefficients* | | |
|---|---|---|---|---|---|---|
| *Model* | | B | Std. Error | Beta | *t* | *Sig.* |
| 1 | (Constant) | 6.762 | 1.362 | | 4.963 | .000 |
| | L1 | .127 | .194 | .104 | .651 | .520 |
| | LOS | .000 | .009 | .009 | .054 | .957 |
| | TimeStudy | .205 | .055 | .603 | 3.734 | .001 |
| | EdDegree | -.188 | .493 | -.061 | -.382 | .705 |

a. Dependent Variable: Sociolinguistic

### Psychological Knowledge Regression Coefficients[a]

| | | *Unstandardized Coefficients* | | *Standardized Coefficients* | | |
|---|---|---|---|---|---|---|
| *Model* | | B | Std. Error | Beta | *t* | *Sig.* |
| 1 | (Constant) | 1.407 | .816 | | 1.723 | .096 |
| | L1 | .186 | .116 | .270 | 1.600 | .121 |
| | LOS | .003 | .005 | .091 | .547 | .589 |
| | TimeStudy | .093 | .033 | .483 | 2.837 | .009 |
| | EdDegree | .020 | .295 | .011 | .068 | .946 |

a. Dependent Variable: Psychological

### Rhetorical Knowledge Regression Coefficients[a]

| | | *Unstandardized Coefficients* | | *Standardized Coefficients* | | |
|---|---|---|---|---|---|---|
| *Model* | | B | Std. Error | Beta | *t* | *Sig.* |
| 1 | (Constant) | 2.016 | .637 | | 3.166 | .004 |
| | L1 | -.018 | .091 | -.035 | -.197 | .845 |
| | LOS | .000 | .004 | -.011 | -.061 | .952 |
| | TimeStudy | .061 | .026 | .428 | 2.369 | .025 |
| | EdDegree | -.130 | .230 | -.101 | -.566 | .576 |

a. Dependent Variable: Rhetorical

# APPENDIX G

## Report Card for Test-Takers

| | |
|---|---|
| **81-100%** | A test taker with this score has reached a high level of pragmatic ability to identify and select language that is appropriate to the context of the situation, considering characteristics of the setting and participants, interactional practices, tone and stance, and genre conventions. |
| **61-80%** | A test taker with this score possesses an adequate level of pragmatic ability to identify and select language that is appropriate to the context of the situation, considering characteristics of the setting and participants, interactional practices, tone and stance, and genre conventions. Further exposure to various settings and language use situations will help |
| **41-60%** | A test taker with this score is in the process of developing pragmatic ability, which is reflected in the inconsistency in their ability to identify and select language that is appropriate to the context of the situation, considering characteristics of the setting and participants, interactional practices, tone and stance, and genre conventions. |
| **21-40%** | A test taker has a limited set of pragmatic resources to identify language that is appropriate to a given situation. |
| **0-20%** | A test taker fails to identify language that is appropriate to a given situation, probably due to lack of exposure and developing language skills. |

Note that lack of time to complete the test limits or invalidates the interpretation of these scores.

| Type of knowledge (Purpura, 2017) | Description for each subcomponent of pragmatic ability | Your score |
|---|---|---|
| **Sociolinguistic Knowledge** | Knowledge of social norms, expectations, and preferences (e.g. distinguishing formal from informal context, evaluating participant roles, etc.). | |
| **Psychological knowledge** | This type of knowledge is related to the communication of emotions, stance, attitudes, and affection. | |
| **Sociocultural knowledge** | This type of knowledge relates to understanding social norms, expectations, and preferences within a particular community (e.g. understanding what may be appropriate in American culture). | |
| **Interactional knowledge** | Knowledge of interactional practices, turn-taking, sequencing, repair (e.g. understanding what the impact of an interruption has on the language use situation). | |
| **Situational knowledge** | This type of knowledge allows a language user to identify aspects of the context that are relevant to the language use situation, e.g. to understand implied meanings that are local to that context. | |
| **Rhetorical knowledge** | This type of knowledge relates to discourse modes and genres, that is, expectations of structuring according to the type of speech event that is in place. | |

Jorge Beltrán is a doctoral student in the Applied Linguistics program at Teachers College, Columbia University (TCCU), specializing in second language assessment. He has taught EFL, ESL, and Spanish in various contexts. He has presented his research at important conferences such as AAAL and LTRC. His research interests include scenario-based assessment, learning-oriented assessment, assessment of speaking ability, construct validation, and performance-based assessment. Correspondence should be sent to him via email: jlb2262@tc.columbia.edu