Understanding the Role of Likeability in the Peer Assessments of University Students’ Oral Presentation Skills: A Latent Trait Approach

Abstract

The purpose of the present study was two-fold: (1) it examined the relationship between peer-rated likeability and peer-rated oral presentation skills of 96 student presenters enrolled in a science communication course, and (2) it investigated the relationship between student raters’ severity in rating presenters’ likeability and their severity in evaluating presenters’ skills. Students delivered an academic presentation and then changed roles to rate their peers’ performance and likeability using an 18-item oral presentation scale and a 10-item likeability questionnaire, respectively. Many-facet Rasch measurement was used to validate the data, and structural equation modeling (SEM) was employed to examine the research questions. At an aggregate level, likeability explained 20% of the variance of the oral presentation ratings and 8% of rater severity. At an item-level, multiple cause-effect relationships were detected, with the likeability items explaining 6% to 30% of the variance in the oral presentation items. Implications of the study are discussed.

Keywords: Likeability; many-facet Rasch measurement; oral presentations; peer assessment; structural equation modeling
Understanding the Role of Likeability in the Peer Assessments of University Students’ Oral Presentation Skills

Likeability is the degree of favourableness of an individual and is due to multiple factors including charisma, confidence, physical appeal, and the ability to build a sense of rapport with people (Reysen, 2005). Studies show that likeability is intimately related to multiple socio-psychological constructs such as social status (Anderson, John, Keltner, & Kring, 2001; Cillessen & Rose, 2005), personality and agreeableness (van der Linden, Scholte, Cillessen, te Nijenhuis, & Segers, 2010), classroom seating and physical distance (van den Berg & Cillessen, 2015), and Internet addiction (Iacovelli & Valenti, 2009). Previous research has also noted that judgements and evaluations can be tainted by the likeability of the evaluated individuals (Carmona, Iyer, & Reckers, 2014).

Despite the importance of likeability in assessment (specifically face to face and oral presentations’ assessment), no previous study has directly investigated its role in language assessment. The mainstream literature has primarily examined the impact of personality and background factors on the assessment of oral proficiency, such as familiarity with peers (O’Sullivan, 2002), interactional features of oral assessments in the First Certificate of English (FCE) speaking test (Dimitrova-Galaczi, 2004), extroversion/introversion (Berry, 2004), assertiveness (Ockey, 2009), interviewer and interview personality and interactions (A. Brown, 2003; Lazaraton, 1996), and linguistic features of oral presentations (Dewaele & Furnham, 2000). Despite the lack of research into likeability in language assessment, a substantial body of literature has yielded evidence supporting its role in performance assessment (Cillessen & Rose, 2005; van der Linden et al., 2010; Lederman, 2012). It has been argued that likeability is an important aspect of individuals’ personality (Lederman, 2012), and personality itself has attracted some attention from language assessment researchers (Ockey, 2009). Nevertheless, Ockey has raised concerns over the relative lack of research in this area, and the present study is a step towards addressing this gap.
One area in which likeability can play a significant part is the assessment of oral presentations. In classroom environments, specifically in low- and medium-stakes assessments, oral presentation skills may be graded partly by peers (Langan, Shuker, Cullen, Penney, Preziosi, & Wheater, 2005). In these environments, the likeability factor can confound peer ratings and, depending on the stakes of the assessment, may affect the validity of the uses and interpretations of the marks (Kane, 2013). For example, Ciarrochi and Heaven (2009) showed that likeability inflates peer assessments in opposite-sex ratings but deflates same-sex evaluations. In other words, likeable students might be over-rated and awarded higher status by their peers. This further indicates that the severity of raters assessing oral presentations can shift due to the likeability of presenters. Similar severity shifts leading to cognitive biases and affect individuals’ impressions and judgements of people or objects have been reported in writing research. For example, Engelhard (1994, 1996, 2012) found that trained raters rating student performance are influenced by such cognitive biases—which he recognizes as the halo effect—thus overestimating students’ proficiency; that is, the presence of the halo effect exerts a sizeable impact on raters’ severity level. Engelhard, however, did not elaborate on the possible causes of such cognitive biases and severity shift in rater performance. The available research suggests that physical attraction and affability could be two primary contributing factors (Ciarrochi & Heaven, 2009; Reysen, 2005), and these potential causes are closely related to likeability (Reysen, 2005).

The current study seeks to examine the relationship between peer-rated likeability of student presenters and peer-rated oral presentation skills. It also explores the relationships between student raters’ severity in evaluating presenters’ likeability and their severity in evaluating presenters’ skills. Likeability is conceived as a multicomponenetal latent trait and measured by Reysen’s (2005) likeability scale, and oral presentation is assessed by a tertiary-level English oral presentation scale (TEOPS) (Aryadoust, 2016). It is postulated that presenters’ likeability may evoke higher peer evaluations of oral presentations. The present study
incorporates the psychometric reliability concept within the framework of many-facet Rasch measurement (MFRM) and structural equation modeling (SEM) to investigate the psychometric quality of TEOPS and the likeability scale.

LITERATURE REVIEW

Behavior and Personality in the Assessment of Oral Proficiency

Research shows that the assessment of oral proficiency is affected by the behavior and personality of raters and test takers (O’Sullivan, 2002). Lazaraton (1996) suggested that interviewees’ oral production may steer and affect assessment by encouraging some interviewers to offer scaffolding and support, which can affect the ratings. According to Lazaraton (1996, p. 167), it is incumbent on researchers to apply a “serious statistical treatment” to determine the impact of such patterns, which can result in unfairness\(^1\) in some cases. Lazaraton’s concerns were voiced in A. Brown’s (2003) study of the speaking test of the International English Language Testing System (IELTS) where interviewers’ differential employment of assessment techniques was a cause of interviewees’ differential performance.

In another study, Ockey (2009) examined the effect of assertiveness (a dimension of extroversion) on the assessment of oral presentations and found that assertive interlocutors were at an advantage when they were paired with non-assertive students, suggesting the ameliorating effects of the proficiency level and non-assertiveness of group members on individual test takers’ ratings. Ockey’s (2004) findings, however, are inconsistent with Berry’s (2004) study in which both extroverted and introverted students taking group oral examinations benefitted from being grouped with extroverts. Ockey conjectured that the observed difference was attributed to the high-stakes context of his study, where the students competed aggressively, a behavior likely penalized by the raters. Ockey’s justification points to the presence of construct-irrelevant variance.

\(^1\) This term is not synonymous with differential item/person functioning or bias in this paper; rather, it is a general terminology and refers to the lack of discriminant validity occasioned or facilitated by construct irrelevant variance.
variance (the amount of error variance in data not explained by the latent variable being measured) in assessments of oral performance, which, in this case, was occasioned by the personality characteristics of the test takers.

In sum, it is plausible that certain personality traits and behaviors of test takers/examiners interact with the assessment environment and exert sizeable influences over student ratings, yielding construct irrelevant variance. Nevertheless, concerns have been raised over the relative lack of research in this area (Ockey, 2009), specifically in peer assessments of oral presentations (Aryadoust, 2016). The present study seeks to address this gap by investigating the relationships between likeability and peer ratings of oral presentations at both the rater level and oral presenter level. Likeability as a factor is highly subjective, and thus flies directly in the face of the need to be objective in assessment. There are many instances in ordinary life where people making important decisions are persuaded by people they “like”—sometimes almost instantaneously. In addition, a great deal of important decisions is made by listening to oral presentations, so, if likeability is an influencer, then student raters may either take advantage of it or fall prey to it.

**Dimensions of Oral Presentations in Academia**

The three-dimension model of oral presentations described by Aryadoust (2016) is adopted in this study. For oral presentations in academic contexts, these dimensions are verbal communication, non-verbal communication, and content and organization.

**Verbal Communication Dimension**

Potentially the most important dimension for oral presentations in academic environments, verbal communication encompasses lexico-grammatical knowledge (Crossley, Clevinger, & Kim, 2014), fluency, and pronunciation and prosodic features (Kang, 2012). Prior research has shown that vocabulary and grammar have a strong influence on the impact of oral presentations (Crossley et al., 2014; Farnsworth, 2013). For example, presenters with larger vocabularies are perceived by audiences as being more capable (Crossley et al., 2014), while those with high grammar knowledge are better able to present their arguments in an organized, logical way.
(Brooks & Swain, 2014; Luoma, 2004) and be universally understood by audiences (Givón, 1995). In contrast, presenters with low grammar knowledge tend to receive poor assessments on graded verbal communication tasks (Crossley et al., 2014). Put broadly, possessing higher lexico-grammatical knowledge helps individuals to be more eloquent, persuasive, engaging, and effective in their oral presentations.

In addition to lexico-grammatical knowledge, effective speakers have a solid grasp of language and the ability to tailor their words to their audience (Jensen & Harris, 1999). While audiences are generally understanding of pauses and other minor disfluencies during the course of a presentation (Taylor, 2011), the attention of the audience and the message of the speech will be lost if the disfluencies are too frequent. Other elements of effective oral presentations include clear enunciation, an appropriate speech delivery rate (Kang, 2012; Kormos & Dénes, 2004; Xi, 2007), and a suitable tone (Pickering, 2004). Poor pronunciation, a rate of speech that is too rapid, and a monotonous tone without proper inflection (Levis & Pickering, 2004) may cause an oral presentation to fail.

**Non-Verbal Communication Dimension**

Non-verbal communication is also an important element of oral presentations. This is a broad dimension and encompasses eye contact (Palmer, 2011), facial expressions, body language (Gui, 2012), hand and body movements (Gui, 2012), clothing (Berkun, 2010), poise, time management, and confidence (Behnke & Sawyer, 2004). Audio-visual aids, such as PowerPoint slides, can also be included under the non-verbal communication domain. When such non-verbal communication elements are used effectively, they can help engage the audience and illustrate the content of the speech (Brock & Joglekar, 2011; Cyphert, 2004). Research has shown that appropriate eye contact and facial expressions are essential for helping speakers maintain the attention of the audience for the duration of a speech, and that the lack of such non-verbal communication elements reduces the effectiveness of a speech (Sweeney, 1998).

**Content and Organization Dimension**
This dimension of oral presentations concerns the topic or the subject matter presented (content) as well as the cohesion and coherence of the presentation (organization). With regard to the content, the first major component of this dimension, a presenter who is familiar with the audience’s level of knowledge of the subject matter—as well as their culture, insights, and attitudes—will be better understood by the audience, and thus would deliver a more effective speech (Tsui & Fullilove, 1998).

For any presentation with structured content—especially academic presentations—this third dimension is essential as it allows for smooth flow and effective transitions (Brooks & Swain, 2014). The use of discourse markers allows the presenter to progress from one topic to the next, and also aids the audience in keeping up with the presenter by alerting them when the topic is about to change (Hansen & Jensen, 1994). A lack of discourse markers has been shown to impede an audience’s ability to thoroughly understand a speech, as shown in the context of listening comprehension tests as well as academic lectures (Hansen & Jensen, 1994; Tauroza, 1995).

The content and organization dimension has several commonalities with the pragmatic dimension of Purpura’s (2004) model of grammar—Purpura’s model also comprises grammatical forms and meaning. The content and organization dimension comprises the awareness and application of the genre (rhetorical meaning) which is represented by suitable introductions and conclusions alongside the rhetorical development of the arguments. In addition, it takes into account the development of clear contexts for the ideas and the provision of a sound analysis in a manner suitable for the audience (Aryadoust, 2015, 2016) which, in Purpura’s model, is subsumed under the contextualized meaning or the mutual understanding of the topic. Finally, the content and organization dimension engages the use of cohesive or transitional devices and operationalizes the sociolinguistic and sociocultural dimensions of pragmatic meaning by taking into account presenters’ politeness, formality, and the meaning associated with the academic culture in the context of the study (Purpura, 2004).
Likeability Research

In everyday life, likeability refers to the quality of being amiable and pleasant. As a personality construct, likeability is a quality of individuals who possess perceived pleasant characteristics, an attractive physique (Dion, Berscheid, & Walster, 1972; Verhulst, Lodge, Lavine, 2010), and affable manners (Reysen, 2005). Studies show a few distinctive characteristics that can be observed among likeable people: they project a positive image of themselves, establish enduring friendships (Cillessen & Rose, 2005), show helpful social behaviour and low levels of aggressiveness (van der Linden et al., 2010), and are trusted more than ordinary individuals (Lederman, 2012). Many people also deem individuals who are similar to themselves as likeable, and this, in turn, constructs a sense of trust toward the likeable individuals (Lederman, 2012). Lederman suggests that non-verbal communication aligned with a verbally communicated message has an instrumental role in promoting likeability.

There is a growing body of literature that recognizes the importance of likeability in (peer) evaluations performed in social and educational settings. Launay and Dunbar (2015) demonstrated that small groups allocate higher likeability ratings to strangers who share certain identifying traits with them, whereas large groups, where the “sense of exclusiveness” is not as marked, assign lower likeability ratings to strangers. Similarly, Brown and Knight (1994) wrote that the participants in their study provided higher ratings to their friends, their group mates, and dominant students. These findings suggest that many inferences drawn from peer assessment scores may not be accurate enough to reflect the ability of the assessed students.

Several studies have documented the association between likeability and personality of individuals in academic settings. Van der Linden et al. (2010) found that peer-evaluated likeability of students is associated with elements of students’ self-image, such as the extent to which they perceive themselves as extrovert, agreeable, and emotionally stable. Carmona et al.’s (2014) study also supports the role of likeability in performance evaluations in organizations across collectivist and individualistic societies. Comparing American and Spanish students...
enrolled in a Master of Business Administration (MBA) program, Carmona et al. found that likeability, along with similarity to self, had a greater impact on the performance evaluations performed by Spanish participants, whereas American students were primarily influenced by financial incentives. This study supports the mediating role of culture in influencing peer assessments performed by tertiary-level students and suggests that overgeneralizing findings across different cultural contexts results in imposing a set of different cultural paradigms on essentially different cultures. Overall, it may be said that likeability affects judgements in different ways that are determined by the cultural and psychological features of individuals in society.

Likeability can further influence class seating and physical distance. Van den Berg and Cillessen (2015) recently showed that students around the age of 12 sit near likeable peers and that students sitting in the middle rows were granted higher likeability scores by their peers. Toward the end of the school year, however, the association between distance and likeability gradually dwindled, showing that distance can predict likeability only at the stage when friendships are developing; after this period, when students have established friendship bonds, likeability and distance gradually disassociate.

Although the role of likeability in assessments is rather under-researched, the aforementioned survey of the literature suggests that there is a role for likeability in peer assessments of oral presentations, which is the focus of the present study. In the present study, likeability is conceptualized as a multicomponential construct comprising warmth, friendliness, similarity to self, approachability, knowledge, and physical attraction (Reysen, 2005). Reysen’s questionnaire comprises 10 items (see Appendix 1) with a Likert-type scale from 1 (very strongly disagree) to 7 (very strongly agree)\(^2\). The scale was slightly modified for the purposes of the current study.

\(^2\) The original likeability scale comprises 11 items, but two of the items were merged to make it suitable for this study.
Research Questions

As previously discussed, the role of likeability as a confounding factor in individuals’ judgements can be significant, particularly for oral presentations. Accordingly, the present study addresses the following research questions (RQs):

(1) Does student presenters’ peer-rated likeability predict the peer-rated oral presentation skills?

(2) Does student raters’ severity in evaluating presenters’ likeability predict their severity in rating presenters’ skills?

If likeability is shown to have a significant effect on the peer assessments of oral presentations, it would be considered a source of potential unfairness and construct irrelevant variance (Messick, 1989). Construct irrelevant variance is yielded by systematic sources of error and distorts ratings of student language performance, thereby rendering the uses and interpretations of the ratings invalid and imprecise. In the present study, construct irrelevant variance is treated as tantamount to the high association between peer ratings and their likeability—leading to the lack of discriminant validity (Campbell & Fiske, 1959)—and it is examined using structural equation modeling (SEM).

Whereas previous studies have largely investigated the association of holistic likeability ratings with other psychological constructs, the present study seeks to examine the relationships between the constituents of likeability and those of the TEOPS dimensions. The great majority of previous studies—which inform RQs 1 and 2—have used scale variables (aggregate-level variables) to explore likeability associations with other psycho-social constructs (e.g., Carmona et al., 2014; Van den Berg & Cillessen, 2015; Van der Linden et al., 2010), but there is scant research on the functionality of the constituents of likeability. While using aggregate-level variables is pragmatic and facilitative, this procedure has proved less useful for displaying and using all of the dimensions that cause variance (Widaman, 2002). Therefore, rather than merely...
using aggregate-level variables, item-level variables are also used in the present study to examine
the finer-grained relationships among the constituents of likeability and TEOPS. Accordingly,
the third research question is formulated as follows:

(3) Can student presenters’ peer-assigned ratings on the items of likeability predict their
ratings on the items of verbal communication, non-verbal communication, and content
and organization?

METHODOLOGY

Participants

The data was collected in a Singaporean university course entitled Science Communication
offered by the Faculty of Science. The participants were 96 freshmen (female=34; male=62)
from three classes and aged between 18 and 21 years. The academic majors of the participants
were life science ($n=44$), statistics ($n=25$), chemistry ($n=20$), and physics ($n=7$). This sample
size satisfies the requirements for reliable MFRM analysis, which is at least 30 individuals (M
Linacre, personal communication, December 8, 2015; see also Linacre, 1994). It also satisfies
the sample size requirements for SEM, which fall between the person-to-item ratios of 5 to 1
(Bentler & Chou, 1987) and 20 to 1 (Tanaka, 1987)—see Wolf, Harrington, Clark, & Miller,
2013, for further discussion).

Instruments

Tertiary-Level English Oral Presentation Scale (TEOPS)

The instrument used to evaluate the presentations was TEOPS, an 18-item evaluation form based
on recommendations and review of the oral presentation research mentioned above (see
Aryadoust, 2015, 2016, for further information). TEOPS has a three-category scoring scale
(1=fair/needs improvement; 2=good/passes; and 3=excellent/outstanding) and contains three
dimensions: content and organization (4 items), non-verbal communication (6 items), and verbal
communication skills (8 items) (see Appendix 1), which are further discussed below. TEOPS
has been shown to be suitable for self-, peer, and teacher assessment of oral presentations and the provision of feedback to student presenters (Aryadoust, 2015, 2016).

**Likeability Scale**

The likeability scale developed and validated by Reysen (2005) was used to measure the likeability of student presenters. I chose this instrument over other measures because of its reported high reliability and because the other available instruments contained components unfit for the context of the current study.

The instrument conceptualizes *likeability* as a construct including multiple constituents such as friendliness, similarity to self, approachability, knowledge, and physical attraction (see Appendix 2). The original scale contains 11 items; however, for the context of the present study, items 9 and 10 (which ask whether the person would make a good roommate or co-worker) were merged, since co-working in a business sense is irrelevant to undergraduate students’ academic life. The instrument comprises a seven-point Likert-type scale ranging from *very strongly disagree* (1) to *very strongly agree* (7).

**Procedures**

During the first half of the semester, the students received guidance from the researcher on how to rate oral presentations by their peers. The overall process included practice brief rating sessions, norming, and presentation rehearsals. First, the students were taught the characteristics of effective oral presentations in detail, such as having an effective introduction, using connector and filler words, and having closing remarks. They viewed several speeches available on YouTube of effective presenters from *The World Science Festival*, and then compared these successful speeches with inferior speeches using an evaluation form. The students also watched and assessed other YouTube presentations of various lengths as homework.

Next, several brief norming sessions were conducted by the teacher in each class to train the students on how analytical scales can be used to identify different aspects of oral presentations and assess them in an analytical way, which is more challenging when a holistic
scale is used. In addition, a few students gave presentations that were rated by their peers. These peer ratings were analyzed and compared with the ratings assigned by the researcher. Feedback regarding the presentations, peer ratings, and analysis of peer ratings were given to the students.

Lastly, each student was required to present a prepared 10- to 12-minute speech to the class. The topic of the speech was any of the five science books used in this communications module (see Appendix 3 for book titles), and the majority of students chose a topic related to their academic major. The students played both roles, presenting to the class and being rated by their peers, as well as listening to presentations by their peers and then rating the presentations. Students were only required to evaluate presentations by students in the same class. The researcher rated all student presenters from all classes, which helped link the data (M Linacre, personal communication, March 13, 2014). Depending on the number of students in a class, each student presenter was rated by 10 to 15 student raters and the teacher.

**DATA ANALYSIS**

Multiple data analyses were carried out in the present study, which are summarized in Table 1. Overall, two major groups of many-facet Rasch measurement (MFRM; see Equation 1 and 2 below) alongside two major groups of structural equation models (see Figure 1 and 2 below) were estimated to answer the three aforementioned research questions. These analyses are discussed in more details as follows: first the scale-level MFRM is presented, followed by the item-level MFRM. Next, the scale-level structural equation models are demonstrated, and finally the item-level structural equation models are discussed.

**Scale-Level Many-Facet Rasch Measurement**

An important desideratum of measurement and assessment is psychometric quality. In this study, MFRM was estimated using the FACETS computer package (Linacre, 2015a) to check the reliability of the data obtained using TEOPS and the likeability scale before answering the research questions. In this section, the MFRM analysis performed separately on the two psychometric instruments is discussed.
Within MFRM, the chance of a student presenter being awarded a specific rating on each item on the list is a function of their ability, raters’ severity, item difficulty, and scoring rubrics (Linacre, 1989). The mathematical expression of the model for measuring oral presentations in the present study is:

$$\log \frac{p_{nijk}}{p_{nijk-1}} = B_n - D_i - C_j - E_k$$

(1)

where

- $P_{nijk}$ is the probability that student presenter $n$ is awarded a score of $k$ on item $i$ by rater $j$;
- $P_{nijk-1}$ is the probability that student presenter $n$ is awarded a score of $k$ on item $i$ by rater $j$;
- $B_n$ is the presentation ability (or likeability) of student presenter $n$;
- $D_i$ is the difficulty level of item $i$;
- $C_j$ is the severity level of rater $j$;
- $E_k$ is the difficulty of the threshold from $k-1$ to $k$ on the scale unique to item $i$.

Likewise, the likeability of the presenters was measured using a discrete MFRM analysis comprising presenter likeability ($B_n$), the difficulty level of the likeability components/items ($D_i$), the severity level of rater students in rating presenters’ likeability ($C_j$), and the difficulty level of the thresholds ($E_k$). The output of these analyses were then transferred to AMOS—the structural equation modeling software (Arbuckle, 2014)—to answer research questions 1 and 2 (see Scale-Level Structural Equation Modeling below).

Each facet (e.g., the ability of each student presenter) is signified by an independent parameter with an interval scale called logits (log-odd units). The standard error of measurement is predicted for each parameter, indicating how accurate the estimations are. MFRM analysis generates fair average measures by adjusting the estimated parameters for influences from other facets as well as the linearized presenter ability (presenters’ skill) parameters (Linacre, 2015b).
The data was examined for any possible misrepresentation of the measurements and facets using infit and outfit mean square (MNSQ) statistics (Bond & Fox, 2015). Infit MNSQ is an information-weighted index sensitive to the data patterns of inliers or the student ratings on the items that are close to the ability level of the student. Outfit MNSQ, on the other hand, is sensitive to outliers or the student ratings on the items far away from the ability level of the student. For this study, the productive fit range proposed by Bond and Fox (2015) was applied, which treats values greater than 1.4 as underfitting (misfitting) and values below 0.6 as overfitting.

MFRM reliability statistics – the proportion of true variance to observed variance, which shows the extent to which the elements in each facet spread – were estimated for all facets. The MFRM reliability index ranges from zero to one, with higher values indicating a greater diversity of the elements of each facet and a lower error of measurement (Wright, 1996). Therefore, it is ideal to achieve high student presenter and item reliability coefficients but lower rater reliability coefficients because low rater reliability coefficients point to high degrees of agreement between different raters (Wright & Masters, 1996). Reliability is also expressed as separation (“the spread of the [measures] relative to their precision”) (Linacre, 2015a), which assigns a numerical value to the number of statistically separate strata of performances and ranges from zero to $+\infty$ (Bond & Fox, 2015).

**Item-Level Many-Facet Rasch Measurement**

To prepare the data for answering research question 3, further MFRM analyses were performed. The mathematical expression of the analyses is:

$$\log \frac{p_{nijk}}{p_{nijk-1}} = B_n - C_j - E_k$$

(2)

It should be noted that this mathematical expression is similar to equation 1, except that it does not factor in the $D_i$ parameter (the difficulty level of items)—which makes it analogous
to a rating scale model. In addition, as demonstrated in Table 1, this equation was used to estimate the presenters’ measures per each likeability item (i.e., since only one item would enter the model at a time, there was no need to factor in the item). Next, the model was used to compute the presenters’ measures per each TEOPS item. Likewise, the model was separately applied to estimate the raters’ severity measures per each likeability item; and finally, it was applied to compute the raters’ severity measures per each TEOPS item. The variables generated in this round of MFRM were all linearized (in logits) and were later transferred to the SEM software (AMOS) to answer the third research question (see Item-Level Structural Equation Modeling below).

Normality Check

Whereas Rasch measurement is not based on the normality assumption, the SEM with maximum likelihood method of parameter estimation, which was used in this study, assumes normality. Initially, multivariate normality was evaluated by estimating Mardia’s coefficient of multivariate kurtosis and its critical ratio (CR). Multivariate normality is supported when this ratio is smaller than 1.96 \((p > .05)\) (Mardia, 1970). This analysis was performed for both student presenters and student raters using AMOS, Version 22 (Arbuckle, 2014). In addition to Mardia’s coefficient, this analysis identified the data points farthest from the centroid, also known as the Mahalanobis distance. Overall, seven student raters and two student presenters fell far away from the centroid under the hypothesis of normality; these individuals were deleted, thus rendering the data normally distributed (Mardia’s coefficients < 1.96).

Next, adherence to univariate normality was examined using skewness (an index of symmetry relative to a normal distribution) and kurtosis (an index of peakedness relative to a normal distribution). The skewness and kurtosis values of all variables fell between -2 and +2, indicating that the data satisfied the univariate normality assumption (George & Mallery, 2010). These preliminary data checks prepared the data for the multivariate statistical analyses discussed below.
Scale-Level Structural Equation Modeling (SEM)

Bond and Fox’s (2015) Rasch measurement-based method for SEM was followed in this study. This method linearizes students’ ratings on each item (in log-odd units or logits) and then integrates the linearized measures into SEM. It also identifies unexpected student ratings and helps purifying data before using them in the SEM analysis (Bond & Fox, 2015). The first round of scale-level SEM analysis comprised Model 1a and 1b and was performed to explore the relationships between student presenters’ oral presentation ratings (their linearized measures estimated by MFRM on the three TEOPS dimensions) and their linearized likeability measures (Research Question 1; see Figure 1). The second round of scale-level SEM analysis comprised Model 2a and 2b and investigated whether the severity level of peer raters in rating presenters’ skills would be influenced by their severity in rating these presenters’ likeability (Research Question 2; see Figure 1). The following fit statistics were calculated:

1. \( \chi^2 \) (chi-square): a statistic examining discrepancies of the posited model and the data. Non-significant \( \chi^2 \) indices \( (p > 0.05) \) suggest the fit of the model to the data. Nevertheless, \( \chi^2 \) is highly dependent on the sample size and might return false positives (i.e., to be statistically significant) when the sample size is too small or large. Therefore, it is normed by dividing it by the degrees of freedom (\( \chi^2/df \)), where values < 3 suggest good fit.

2. Root mean square error of approximation (RMSEA): an index correcting for the sensitivity of \( \chi^2 \) to sample size. Values below 0.06 indicate good fit, but if the degrees of freedom are low the RMSEA “too often falsely indicates a poor fitting model” (Kenny, Kaniskan, & McCoach, 2015, p. 486) and, therefore, other fit indices should also be considered.

3. Comparative fit index (CFI): an incremental fit index with values greater than 0.90 indicating reasonable fit. CFI is not affected much by sample size, and is thus a reliable index (Kenny, 2014).
(4) Tucker Lewis index (TLI): another incremental fit index with values greater than 0.90 indicating reasonable fit. TLI is not affected much by sample size (Kenny, 2014).

(5) Consistent Akaike Information Criterion (CAIC): used to compare competing models, i.e., Model 1a vs. 1b and Model 2a vs. 2b. Models with a lower CAIC index are optimal.

The scale-level SEM analyses computed the unstandardized and standardized regression coefficients per path; the former index estimates the increment of the dependent (endogenous) variable if the independent (exogenous) variables increase/decrease by one raw unit. The latter index shows the amount of increment in the dependent variable in standard deviation (SD) units if the independent variables increase/decrease by one SD (Schumacker & Lomax, 2010). Finally, to evaluate the proportion of the variance in the dependent variables explained by the independent variable, squared multiple correlations were estimated.

**Item-Level Structural Equation Modeling**

Dependent variables comprised the linearized presenters’ skill measures on the 18 variables measured by the TEOPS dimensions (see *Item-Level Many-Facet Rasch Measurement* above for further information). Presenters’ linearized measures per likeability item were also estimated separately, yielding 10 independent variables. Subsequently, 18 discrete higher-order SEM models (Schumacker & Lomax, 2010) were generated to predict each TEOPS item based on the 10 independent variables. The models comprised one common measurement model (left-hand side of the diagram in Figure 1) with student presenters’ ratings on 10 likeability items. The two-headed arrows indicate the covariance and improve the fit of the model while capturing dependencies between similar items (Duncan, Duncan, & Strychker, 2006). The second part (right-hand side of the diagram in Figure 1) comprised a lower-order structural model, where the 18 TEOPS items were regressed on likeability items. Following McDonald and Ringo (2002), I deleted the non-significant paths ($p < 0.05$) to arrive at an optimal solution.
Except CAIC, which is used for model comparison, all aforementioned fit indices were estimated to examine the fit of the models.

**RESULTS**

**Many-Facet Rasch Measurement (MFRM)**

Two separate rounds of MFRM analysis were performed: round one (represented by Equation 1 above) included three discrete MFRM analyses of the three dimensions of the TEOPS and the second round (represented by Equation 2 above) comprised the MFRM analysis of the likeability scale. Overall, the spread of distribution of the student presenter ability ratings across the three TEOPS dimensions was invariably higher than were the rater severity scores. The likeability ratings also had a rather large spread for both presenters and raters.

As shown in Table 2, the infit and outfit MNSQ indices of all items tapping the constituents of the TEOPS dimensions indicate sufficient fit to the model, as they fall between 0.6 and 1.4. On the other hand, item 4 under likeability, which is the most difficult item, has a slight misfit; because the item is lowly endorsed by the majority of students, a small number of students endorsing it highly could have caused the misfitting pattern (Linacre, 2003). In addition, the error of measurement is fairly low in estimating the difficulty level of the items, suggesting high precision in the difficulty and fair average measures.

The items reliability indices of the content and organization, verbal communication, and non-verbal communication were .94 (separation = 3.98 levels; $\chi^2 = 127.6$, $p < 0.05$), .98 (separation = 6.81 levels; $\chi^2 = 188.4$, $p < 0.05$), and .97 (separation = 5.42 levels; $\chi^2 = 179.4$, $p < 0.05$), respectively. The reliability and separation indices of the likeability scale were 1.00 and 15.95, respectively. The high reliability indices show that if the scales are used by a different group of students derived from the same population, the item difficulty indices will most likely be reproducible (Linacre, 2015a).

As with the item psychometric qualities, the student presenters all fitted MFRM with infit and outfit MNSQ values between .60 and 1.40. In addition, the reliability (separation) statistics
ranged between .81 (2.09) for non-verbal communication and .98 (6.80) for likeability. This provides evidence that the scales successfully distinguished between two to six strata of ability.

The MFRM analyses reported above computed (1) the student presenters’ ability (in logits), (2) the student raters’ severity in rating the presenters’ skills (in logits) across the three dimensions of TEOPS, (3) the student presenters’ likeability (in logits), (4) and the student raters’ severity in rating the presenters’ likeability (in logits). These variables were then transferred to AMOS (the SEM software) to answer research questions 1 and 2.

**Research Questions 1 and 2: Scale Level SEM Analysis**

To answer the first and second research questions, the linearized oral proficiency and severity ratings were incorporated in four rounds of SEM analysis. As demonstrated in Figure 2, Model 1a represents the overall effect of student presenters’ likeability on their peer-rated oral presentations, whereas Model 1b examines the effect of likeability on each oral presentation dimension separately. Similarly, Model 2a for raters examines the overall effect of likeability, whereas Model 2b estimates separate paths. Models 1a and 1b are rivals and so are Models 2a and 2b.

The observed variables in all models are represented by rectangles, the latent variables by large circles, and the errors of measurement and disturbances by small circles. (Disturbances refer to the amount of variance not explained in oral presentations in Model 1a and 2a). The two-headed arrows represent the covariance between the errors of measurement and have been added to optimize the fit (Duncan et al., 2006). In all models, one of the paths is fixed for parameter estimation.

In Model 1a, students’ oral presentation ratings are indicated (measured) by the three dimensions of TEOPS. This model is more parsimonious than other competing models presented earlier (d.f. = 2) and lends itself to the constituent structure of the data, with a much better fit ($\chi^2=2.682; \chi^2/d.f.=1.341; \text{RMSEA}=0.060; \text{CFI}=.994; \text{TLI}=.983; \text{CAIC}=47.197$) than Model 1b (see Table 3). In Model 1a, the standardized coefficients placed before the slash adjacent to each
one-headed arrow are significantly high \( p < 0.01 \) and represent the effect size of each observed variable—the unstandardized coefficients are presented after the slash. Likeability of the presenters exerts a substantive and significant impact on oral presentation ratings: when the likeability of the presenter increases by one SD unit, the overall rating of their oral presentation increases by 0.40 SD units. Finally, although Model 1b \( (\chi^2=14.588; \chi^2/\text{d.f.}=14.588; \text{RMSEA}=0.382; \text{CFI}=0.883; \text{TLI}=0.297; \text{CAIC}=64.477) \) captures significant associations between likeability and oral presentation dimensions \( p < 0.01 \), it has a suboptimal fit. Overall, the optimal fit of Model 1a suggests that the effect of likeability permeates oral presentation ratings, and the suboptimal fit of Model 1b indicates that this effect cannot be separated at a subscale level where content and organization and (non-)verbal communication are discrete variables.

Similarly, Model 2a \( (\chi^2=3.026; \chi^2/\text{d.f.}=1.513; \text{RMSEA}=0.076; \text{CFI}=0.984; \text{TLI}=0.953; \text{CAIC}=47.025) \) outperformed Model 2b \( (\chi^2=19.542; \chi^2/\text{d.f.}=19.542; \text{RMSEA}=0.456; \text{CFI}=0.719; \text{TLI}=0.687; \text{CAIC}=69.040) \), indicating that the effect of severity in rating likeability permeates peer raters’ severity in rating oral presentation ratings (Model 2a), and that this effect cannot be separated at a subscale level with discrete variables (Model 2b). The RMSEA in the models is inflated due to the low degrees of freedom and the other fit indices should be given priority to evaluate the fit (Kenny et al., 2015).

The evidence yielded against the effect of likeability at a subscale level does not nullify the item-level relationships between likeability and oral presentation items. The following sections will examine these relationships.

**Research Question 3**

Table 4 presents the fit statistics as well as the (un)standardized regression weights of the significant predictors and the squared multiple correlations per model. The measurement model (illustrated earlier in Figure 1) is presented in the first row followed by eight models for content organization, four models for verbal communication, and six models for non-verbal communication. The CFI and TLI of all models indicate very good fit, although the RMSEA
indices are inflated, possibly due to the number of parameters (Kenny et al., 2015). The squared multiple correlations range between .062 (NVC_6 or item 6 tapping into the non-verbal communication dimension) and .300 (VC_3 or item 3 tapping into the verbal communication dimension), meaning that the linearized likeability ratings of student presenters can explain 9% to 30% of the variance in their componential ratings. Overall, approximately 20% or more of the variance in half of the items of oral presentations is explained by likeability ratings.

Three discernible patterns emerge from the (un)standardized regression weight columns. First, the likeability items have varying degrees of predictive power in explaining variance in the 18 oral presentation items. For example, being friendly, approachable, knowledgeable, physically attractive, and a potential roommate are five likeability items explaining 30% of variance in the clarity and accuracy of pronunciation (VC_3), while being knowledgeable predicts 6% of variance in presenting effective audio/visual aids (NVC_6). Second, the impact of some likeability items on student presenters’ ratings is negative (bold print variables). For example, when being approachable increases by one SD, the ratings on interesting opening to capture attention (CO_1 or item 1 tapping into the content and organization dimension) decrease by .653 SDs, which is considerable. Interestingly, except for the three cases highlighted in grey, all likeability items with negative regression weights maintain their negative direction across the oral presentation items. Finally, some likeability items, such as friendliness and being knowledgeable, were able to explain varying shares of variance of multiple TEOPS items, whereas others, such as friendship and warmth, had lower explanatory power. These findings are discussed in further detail below.

**DISCUSSION**

The primary objective of the present study was to examine whether peer assessments of oral presentations are influenced by presenters’ likeability. The raw data were subjected to MFRM analysis and evidence for the fit of the data to the model was garnered, although item 4 of the likeability scale (similarity to self) had a slightly inflated fit index. One possible reason might be
the vagueness of the phraseology of this item. Similarity, as a broad concept, refers to the state of resemblance between two objects or individuals and, as such, it can be interpreted in a number of ways. While some might construe it as physical, behavioural, or personal similarity, others might interpret it as similarity in, for example, presentation style, academic achievement, or other related concepts. The research questions (RQs) are discussed individually below.

**RQ1** The SEM analysis with an oral presentation latent variable (Model 1a) had a better fit than the path model with observed variables (Model 1b), indicating the relationship between the three dimensions of oral presentations measured by TEOPS, and thus the need for an overarching latent variable causing them (Hair, Black, Babin, Anderson, & Tatham, 2006; Schumacker & Lomax, 2010;). Using this latent variable also helped mediate the impact of likeability on the three dimensions of oral presentations. It was found that student presenters’ likeability ratings exerted a holistic and significant impact on oral presentation ratings, predicting approximately 20% of their variance; therefore, the constituents of oral presentations are intimately interconnected and explicable by reference to an overarching latent variable, while 20% of their variance is due to student presenters’ likeability.

The findings of the study are in agreement with Lederman (2012) who argued for a strong relationship between likeability and oral communication skills in everyday life. It is also consistent with work by Brown and Knight (1994), Carmona et al. (2014), and Launay and Dunbar (2015) in that the items of likeability, such as familiarity with the presenter, can inflate the ratings assigned by raters. This suggests that likeability would play a part in the peer assessment of oral presentations. It further implies that other personality attributes could similarly exert an impact on performance assessments, which can (hopefully) become a new line of research integrating psychology and language assessment.

The observed relationships between likeability and oral presentations—while preliminary—have important implications for research. Methodologically, it is recommended that future studies on this or similar topics should apply SEM with latent variables in order to
preserve the variance caused by all dimensions and to reduce the potential collinearity between the variables, which can affect linear regression models (Hair et al., 2006).

The other implication is substantive and pertains to the association of oral presentations and likeability. The finding that likeability affects oral presentation ratings indicates potential unfairness in performance assessments. The presence of certain patterns leading to potential unfairness in interviews and oral presentations has been reported by Lazaraton (1996, p. 166), who expressed uncertainty as to how such patterns should be treated in the IELTS speaking test: “should they be tolerated? Prohibited? Encouraged? And are such prescriptions desirable?” Lazaraton’s study did not find causal relationships between interviewees’ performances and raters’ behaviours (she used qualitative methods), whereas the current study discovered cause-effect relationships through the application of SEM and is therefore in a position to provide preliminary responses to Lazaraton’s questions. Any unfairness gleaned from likeability is construct irrelevant variance and undesirable, and thus must be mitigated through training and moderation. Some practitioners might argue that in formative and low-stakes assessments performed by peers this impact is of low concern, and therefore resources should not be allocated to training students. However, it can be argued that even in such environments—specifically when students attach high importance to peer ratings, like in the present study—receiving unfair ratings from peers may demotivate and demoralize students. It could even lead students to make erroneous conclusions about specific aspects of their performance and adversely impact learning outcomes.

To establish student-centred learning environments and render students accountable for their learning by engaging them in assessments, we need to tease out potential sources of construct irrelevant variance and unfairness in peer assessments and attempt to enlighten and train students to pre-empt these undesirable impacts. To maximise the benefits of peer assessment, this is what is necessary. Safeguarding peer-assessments against potential construct irrelevant variance is well worth our time, resources, and endeavours because, by training our
students to be fair in their judgements, we will prepare them to take responsibility in their academic lives and future careers while raising their awareness about potential unfairness in every judgement.

**RQ2** The SEM analysis results showed that the severity of the student raters in rating likeability explains approximately 8% of variance in severity in rating content and organization. Put differently, when the severity of the student raters in rating their peers’ likeability increases by 1 SD, their severity in rating oral presentations increases by 22%. This magnitude is relatively small, suggesting that severity cannot be viewed as a single cognitive module and can vary across different tasks. For example, while a certain student rater might tend to be severe in rating his/her peers’ likeability (would not easily like people), s/he could be quite lenient in rating their presentation skills and vice versa. This finding is partly consistent with previous studies where likeability was associated with raters’ judgement (Cillessen & Rose, 2005; Huesmann et al., 2009; Launay & Dunbar, 2015), but is unique in the sense that it primarily hinges on rater severity—as opposed to previous research where severity across various constructs was not examined. The study is one of the first that quantifies peer rater severity across different constructs to examine their potential relationships.

**RQ3** The findings indicate that item-level analyses can reveal numerous otherwise masked relationships. Nevertheless, since the results were yielded at item levels, they cannot be specifically discussed in light of the literature surveyed earlier, as previous research has mainly focused on the inter-construct relationships.

First, between 20% and 30% of the variance of nine (50%) of 18 total items of oral presentations was explained by an array of likeability items. Each likeability item predicted between three and ten oral presentation items ($p<0.05$). It is important to note the differential role of likeability items in determining presenters’ ratings. For example, friendliness and knowledgeableness play a significant part in predicting the ratings on multiple oral presentation items. Further research needs to be undertaken to investigate the effect of each likeability item.
on the judgements made by trained raters assessing candidates in high-stakes tests. By contrast, other items such as friendship and warmth had lower explanatory power. This could be due to their high correlations, which also means that, statistically, one of these highly associated items would suffice in the analysis. Despite their similarity, however, these items are distinct and necessary in the context of this study as the students had already spent (at minimum) one semester together, forming friendship bonds. As may be expected, not every friend can be truly friendly (item 2), and this might explain why this item was endorsed relatively lowly. Warmth (item 10) is yet another concept concerning the affection and kindness that one might feel from a friend, whereas friendliness is more of a physical manifestation displayed by another individual.

In addition, some likeability items had a negative impact on student presenters’ ratings. For example, CO_1 (opening of presentations) had an inverse cause-effect relationship with being approachable and physically attractive, and CO_2 (item 2 tapping into the content and organization dimension), which pertains to creating context at the outset, also had a negative relationship with being physically attractive. The actual cause of this finding remains to be investigated, but based on my observations in classrooms, there seems to be higher expectations of presenters who appear attractive and approachable. Through informal conversations with some students, I noted that being attractive and approachable would signal to the audience that the presenters will be an effective communicator. It may be that attractive appearance is a major cause for halo effects and cognitive biases (Verhulst et al., 2010). What needs to be further researched is that if raters’ high expectations of attractive presenters are not satisfied, will their leniency/severity shift significantly? Rather than cross-sectional research, like the present study, (quasi-)experimental studies are more suitable to address this question.

On the other hand, being friendly and knowledgeable positively predicted the ratings on CO_1, indicating that high ratings were granted to more knowledgeable and friendly presenters. Whereas the relationship between knowledge and CO_1 is meaningful, the inverse effects of
friendliness and approachability remain a mystery and need to be further examined in the future. If the speculation about the effect of physique and approachableness is accurate, it is expected that rater perceptions and the influence of sources of construct irrelevant variance (approachable and physically attractive) will fluctuate throughout the presentation: while an approachable and physically attractive presenter might be underrated at the outset (CO_1), a different set of likeability items should influence her/his ratings at the conclusion (CO_7). This speculation is supported by the significant effect of warmth and similarity to self on CO_7, suggesting a dynamic nomothetic network of relationships in peer assessments.

However, this hypothesis does not pertain to CO_8 (answering questions), which also has inverse relationships with being approachable and physically attractive. It needs to be borne in mind that this item is relevant only when the presentation is over and the presenter engages in a conversation with the audience. During the data collection and the norming sessions, I noted that some student presenters attempted to compensate for their lack of pertinent response by evading the actual question and showing off their knowledge of a related topic; in addition, some attempted to compensate by showing how approachable and friendly they were, again evading the question; other negative cause-effect relationships can be similarly justified. The mechanisms underlying such potential interactions would be an interesting area for further investigation.

Finally, the vast majority of the significant regression weights in the present study were of weak to medium strength. Two factors might have contributed to this pattern. The first reason is nomological and pertains to the intrinsic relationships between peer assessment of oral presentations and likeability. It may be said that likeability cannot be the only predictor of peer evaluations. It would be useful to generate a nomological network of relationships that accommodates a number of predictors and use a statistical technique to evaluate the posited models. The second possible reason is informed by Van den Berg and Cillessen’s (2015) study:

3 A network of relationships between closely related constructs.
before conducting the study and collecting the data, the students had a great deal of exposure and in-class communication over one academic semester. They were required to interact and collaborate with peers sitting in different locations in class, which led to their positive thinking about each other and a sense of rapport and trust (as reflected in their module evaluation comments). As a result, their ratings of peers’ oral presentation skills were somewhat confounded by their peers’ likeability.

As previously discussed, Carmona et al. (2014) found that likeability had different degrees of impact on ratings given by raters of different cultural backgrounds (Spanish vs American). Future research can investigate possible patterns emerging as a result of the effect of students’ homogeneous/heterogeneous cultural backgrounds on the ratings.

CONCLUSION

Overall, likeability had an impact on ratings assigned by peer raters. Similarly, the severity of peer raters in rating presenter likeability predicted their severity in rating the three dimensions of oral presentations. In addition, several item-level cause-effect relationships were identified. It may be said that peer assessments of oral presentations are affected by presenter likeability, with a weak to medium magnitude, which could influence the validity of the uses and interpretations of the scores. Therefore, caution must be exercised in adopting peer assessments, and the weight of the peer assessment in the final grade (if any) should vary according to the quality of the ratings.

The peer assessments analyzed in this study stem from data collected at the end of the semester when the students had made friendship bonds with each other. The end of the semester might be the right time for collecting peer assessments, provided that the students have been engaged in this type of assessment throughout the semester. It is important to note that likeability remains (or perhaps improves as) a significant factor at the end of the semester. Teachers should remind their students that likeability cannot/should not save a poorly delivered presentation, although in a well-delivered presentation it can help to create an even better impression. In
reliable peer assessments of oral presentations, if the presenter seems pleasant and affable but the presentation is not effective, the raters should not be influenced by the amiability of the presenter.

Finally, these findings open the possibility of likeability being an intervening factor in face-to-face speaking tests/interviews where a test taker might have techniques to impress the assessor beyond his/her verbal skills. The likelihood of such cognitive biases arises from the fact that the assessor and test taker are supposedly not known to each other (especially in high-stakes tests), and there is a high chance that a test taker with a pleasant character will, wittingly or unwittingly, take advantage of his/her first impression and likeability, thereby affecting the ratings. Despite their potential roles in assessment, psychological factors such as likeability have not been given sufficient research attention in performance assessment and further research is needed to investigate this area.

References


Table 1
Summary of Data Analyses

<table>
<thead>
<tr>
<th>Data analysis</th>
<th>Relevant model</th>
<th>Models fitted</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale-level Many-Facet Rasch Measurement</td>
<td>Equation 1</td>
<td>• 3 separate scale-level models for the three dimensions of TEOPS</td>
<td>Each model computes the student presenters’ measures (in logits) which were transferred to</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 1 separate scale-level model for the likeability items</td>
<td>1 Model 1a &amp; Model 1b</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Each model computes the student raters’ severity measures (in logits) which were transferred to</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Model 2a &amp; Model 2b</td>
</tr>
<tr>
<td>Scale-level structural equation modelling</td>
<td>Figure 1</td>
<td>• 4 separate scale-level models (Model 1a, Model 1b, Model 2a, Model 2b)</td>
<td>The models was generated to answer research question 1 (Model 1a &amp; Model 1b) and research question 2 (Model 2a &amp; Model 2b)</td>
</tr>
<tr>
<td>Item-level Many-Facet Rasch Measurement</td>
<td>Equation 2</td>
<td>• 18 separate item-level models, each per one of the 18 TEOPS items</td>
<td>Each model computes student presenters’ ability (in logits) on separate TEOPS items</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 10 separate item-level model, each per one of the 10 likeability items</td>
<td>Each model computes student presenters’ ability (in logits) on separate likeability items</td>
</tr>
<tr>
<td>Item-level structural equation modelling</td>
<td>Figure 2</td>
<td>• 10 separate item-level presenter models</td>
<td>Each model measures the impact of the presenters’ measures of individual likeability items on presenters’ measures per each TEOPS item</td>
</tr>
</tbody>
</table>

Table 2
Psychometric Quality of the three Dimensions of TEOPS and the Likeability Scale

<table>
<thead>
<tr>
<th>Items</th>
<th>Difficulty</th>
<th>Fair average measure</th>
<th>Measure</th>
<th>SE</th>
<th>Infit MNSQ</th>
<th>Outfit MNSQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content &amp; Organization</td>
<td>1.26</td>
<td>2.71</td>
<td>0.31</td>
<td>0.06</td>
<td>1.05</td>
<td>1.06</td>
</tr>
<tr>
<td></td>
<td>2.72</td>
<td>2.77</td>
<td>0.00</td>
<td>0.06</td>
<td>0.98</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>2.8</td>
<td>2.85</td>
<td>-0.50</td>
<td>0.07</td>
<td>0.95</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>2.74</td>
<td>2.78</td>
<td>-0.09</td>
<td>0.07</td>
<td>1.02</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>2.76</td>
<td>2.81</td>
<td>-0.24</td>
<td>0.07</td>
<td>1.05</td>
<td>1.03</td>
</tr>
<tr>
<td></td>
<td>2.68</td>
<td>2.72</td>
<td>0.22</td>
<td>0.06</td>
<td>0.96</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>2.67</td>
<td>2.71</td>
<td>0.30</td>
<td>0.06</td>
<td>1.01</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>2.73</td>
<td>2.77</td>
<td>-0.01</td>
<td>0.06</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Non-Verbal Communication</td>
<td>2.60</td>
<td>2.63</td>
<td>0.36</td>
<td>0.05</td>
<td>1.08</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td>2.74</td>
<td>2.77</td>
<td>-0.3</td>
<td>0.06</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>2.59</td>
<td>2.62</td>
<td>0.39</td>
<td>0.05</td>
<td>0.9</td>
<td>0.91</td>
</tr>
<tr>
<td></td>
<td>2.64</td>
<td>2.68</td>
<td>0.16</td>
<td>0.06</td>
<td>0.94</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>2.71</td>
<td>2.74</td>
<td>-0.13</td>
<td>0.06</td>
<td>1.15</td>
<td>1.08</td>
</tr>
<tr>
<td></td>
<td>2.77</td>
<td>2.8</td>
<td>-0.47</td>
<td>0.07</td>
<td>1.00</td>
<td>0.99</td>
</tr>
</tbody>
</table>

Verbal Communication
<table>
<thead>
<tr>
<th></th>
<th>2.71</th>
<th>2.76</th>
<th>-0.22</th>
<th>0.06</th>
<th>0.97</th>
<th>0.94</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>2.76</td>
<td>2.82</td>
<td>-0.57</td>
<td>0.07</td>
<td>0.95</td>
<td>0.87</td>
</tr>
<tr>
<td>3</td>
<td>2.62</td>
<td>2.67</td>
<td>0.22</td>
<td>0.06</td>
<td>1.04</td>
<td>0.99</td>
</tr>
<tr>
<td>4</td>
<td>2.55</td>
<td>2.58</td>
<td>0.57</td>
<td>0.06</td>
<td>1.05</td>
<td>1.07</td>
</tr>
</tbody>
</table>

**Likeability**

<table>
<thead>
<tr>
<th></th>
<th>5.70</th>
<th>5.76</th>
<th>-0.13</th>
<th>0.04</th>
<th>0.80</th>
<th>0.82</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5.86</td>
<td>5.92</td>
<td>-0.4</td>
<td>0.04</td>
<td>0.81</td>
<td>0.83</td>
</tr>
<tr>
<td>3</td>
<td>5.81</td>
<td>5.87</td>
<td>-0.31</td>
<td>0.04</td>
<td>0.79</td>
<td>0.81</td>
</tr>
<tr>
<td>4</td>
<td>4.64</td>
<td>4.68</td>
<td>1.42</td>
<td>0.03</td>
<td>1.44</td>
<td>1.49</td>
</tr>
<tr>
<td>5</td>
<td>5.66</td>
<td>5.72</td>
<td>-0.06</td>
<td>0.04</td>
<td>0.82</td>
<td>0.84</td>
</tr>
<tr>
<td>6</td>
<td>5.87</td>
<td>5.93</td>
<td>-0.42</td>
<td>0.04</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>7</td>
<td>5.54</td>
<td>5.59</td>
<td>0.14</td>
<td>0.03</td>
<td>1.15</td>
<td>1.17</td>
</tr>
<tr>
<td>8</td>
<td>5.51</td>
<td>5.56</td>
<td>0.19</td>
<td>0.03</td>
<td>0.97</td>
<td>0.99</td>
</tr>
<tr>
<td>9</td>
<td>5.44</td>
<td>5.49</td>
<td>0.3</td>
<td>0.03</td>
<td>1.12</td>
<td>1.13</td>
</tr>
<tr>
<td>10</td>
<td>6.03</td>
<td>6.09</td>
<td>-0.74</td>
<td>0.04</td>
<td>1.15</td>
<td>1.10</td>
</tr>
</tbody>
</table>
Figure 1. Scale-level models for presenters’ proficiency and their rated likeability (Model 1a and 1b); rater severity and their severity in rating their peers’ likeability (Model 2a and 2b).

Note: CO = Content and organization; NVC = Non-verbal communication; VC = verbal communication; St = Student presenters.
Figure 2. The item-level SEM models comprising the measurement model (left) and the complete higher-order SEM model (right). “Likeability Students” is the latent variable indicated (measured) by student presenters’ linearized ratings on the 10 likeability components, which cause linearized ratings on the 18 TEOPS components, thereby generating 18 separate higher-order models.

Note: L_M_1=warm; L_M_2=friendly; L_M_3=likeable; L_M_4=similar; L_M_5=approachable; L_M_6=knowledgeable; L_M_7=physique; L_M_8=advice; L_M_9=roommate; L_M_10=friendship.

e = error of measurement.
<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$\chi^2$/d.f.</th>
<th>RMSEA</th>
<th>CFI</th>
<th>TLI</th>
<th>CAIC</th>
<th>Squared multiple correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1a</td>
<td>2.682</td>
<td>2</td>
<td>1.341</td>
<td>0.060 (0.00 - 0.222)*</td>
<td>.994</td>
<td>.983</td>
<td>47.197</td>
<td>Oral Presentation=.195</td>
</tr>
<tr>
<td>Model 1b</td>
<td>14.588**</td>
<td>1</td>
<td>14.588</td>
<td>0.382 (0.225 - 0.567)</td>
<td>.883</td>
<td>.297</td>
<td>64.477</td>
<td>CO=.189; NVC=.172; VC=.062</td>
</tr>
<tr>
<td>Model 2a</td>
<td>3.026</td>
<td>2</td>
<td>1.513</td>
<td>0.076 (0.00 - 0.237)</td>
<td>.984</td>
<td>.953</td>
<td>47.025</td>
<td>Oral Presentation=.084</td>
</tr>
<tr>
<td>Model 2b</td>
<td>19.542**</td>
<td>1</td>
<td>19.542</td>
<td>0.456 (0.294 - 0.643)</td>
<td>.719</td>
<td>.687</td>
<td>69.040</td>
<td>CO=.097; NVC=.020; VC=.009</td>
</tr>
</tbody>
</table>

*Numbers in brackets are confidence intervals. The lower bound is ideally close to zero and upper bound should also be small. The RMSEA in the models is inflated due to their low degrees of freedom and other fit indices should be given priority to evaluate the fit (Kenny et al., 2015).
### Table 4: Fit Statistics of the Component-Based SEM Models of Student Presenters’ Ratings Assigned by Peer Raters

<table>
<thead>
<tr>
<th>Model</th>
<th>χ²</th>
<th>d.f.</th>
<th>χ²/d.f.</th>
<th>RMSEA</th>
<th>CFI</th>
<th>TLI</th>
<th>Unstandardized regression weights of significant predictors (p &lt; 0.05, 0.01)</th>
<th>Standardized regression weights of significant predictors (p &lt; 0.05, 0.01)</th>
<th>Squared multiple correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measurement model *</td>
<td>84.101**</td>
<td>27</td>
<td>3.115</td>
<td>.149 (.114 - .186)</td>
<td>.929</td>
<td>.957</td>
<td>L_M_2 = .191; L_M_5 = .435; L_M_6 = .542; L_M_7 = .241</td>
<td>L_M_2 = .770; L_M_5 = .653; L_M_6 = .422; L_M_7 = .216</td>
<td>.284</td>
</tr>
<tr>
<td>CO_1</td>
<td>122.950**</td>
<td>34</td>
<td>3.298</td>
<td>.145 (.113 - .177)</td>
<td>.950</td>
<td>.919</td>
<td>L_M_2 = .368; L_M_7 = .339; L_M_9 = .428</td>
<td>L_M_2 = .310; L_M_7 = .395; L_M_9 = .316</td>
<td>.196</td>
</tr>
<tr>
<td>CO_2</td>
<td>106.658**</td>
<td>35</td>
<td>3.047</td>
<td>.147 (.115 - .179)</td>
<td>.947</td>
<td>.917</td>
<td>L_M_2 = .180; L_M_7 = .339; L_M_9 = .428</td>
<td>L_M_2 = .310; L_M_7 = .395; L_M_9 = .316</td>
<td>.196</td>
</tr>
<tr>
<td>CO_3</td>
<td>103.773**</td>
<td>37</td>
<td>2.805</td>
<td>.138 (.107 - .170)</td>
<td>.950</td>
<td>.926</td>
<td>L_M_2 = .203</td>
<td>L_M_2 = .359</td>
<td>.129</td>
</tr>
<tr>
<td>CO_4</td>
<td>105.266**</td>
<td>37</td>
<td>2.845</td>
<td>.139 (.108 - .171)</td>
<td>.949</td>
<td>.924</td>
<td>L_M_2 = .264</td>
<td>L_M_2 = .309</td>
<td>.096</td>
</tr>
<tr>
<td>CO_5</td>
<td>102.225**</td>
<td>35</td>
<td>3.007</td>
<td>.145 (.113 - .178)</td>
<td>.949</td>
<td>.918</td>
<td>L_M_2 = .414; L_M_6 = .329; L_M_8 = .541; L_M_9 = .544</td>
<td>L_M_2 = .463; L_M_6 = .352; L_M_8 = .523; L_M_9 = .425</td>
<td>.162</td>
</tr>
<tr>
<td>CO_6</td>
<td>104.459**</td>
<td>37</td>
<td>2.823</td>
<td>.139 (.108 - .170)</td>
<td>.950</td>
<td>.925</td>
<td>L_M_2 = .549; L_M_4 = .197</td>
<td>L_M_2 = .553; L_M_5 = .594; L_M_7 = .353; L_M_10 = .632</td>
<td>.265</td>
</tr>
<tr>
<td>CO_7</td>
<td>104.497**</td>
<td>36</td>
<td>2.903</td>
<td>.142 (.110 - .174)</td>
<td>.949</td>
<td>.922</td>
<td>L_M_2 = .676; L_M_5 = .744; L_M_7 = .312; L_M_10 = .604</td>
<td>L_M_2 = .553; L_M_5 = .594; L_M_7 = .353; L_M_10 = .632</td>
<td>.265</td>
</tr>
<tr>
<td>CO_8</td>
<td>100.846**</td>
<td>34</td>
<td>2.967</td>
<td>.144 (.112 - .177)</td>
<td>.950</td>
<td>.920</td>
<td>L_M_2 = .669; L_M_2 = .109; L_M_5 = .854; L_M_7 = .398; L_M_8 = .560; L_M_9 = .777</td>
<td>L_M_2 = .534; L_M_2 = .819; L_M_5 = .624; L_M_7 = .412; L_M_8 = .456; L_M_9 = .512</td>
<td>.228</td>
</tr>
<tr>
<td>VC_1</td>
<td>100.859**</td>
<td>32</td>
<td>3.152</td>
<td>.151 (.118 - .184)</td>
<td>.949</td>
<td>.912</td>
<td>L_M_2 = .676; L_M_5 = .744; L_M_7 = .312; L_M_10 = .604</td>
<td>L_M_2 = .534; L_M_2 = .819; L_M_5 = .624; L_M_7 = .412; L_M_8 = .456; L_M_9 = .512</td>
<td>.228</td>
</tr>
<tr>
<td>VC_2</td>
<td>108.007**</td>
<td>37</td>
<td>2.919</td>
<td>.142 (.111 - .174)</td>
<td>.947</td>
<td>.922</td>
<td>L_M_2 = .676; L_M_5 = .744; L_M_7 = .312; L_M_10 = .604</td>
<td>L_M_2 = .534; L_M_2 = .819; L_M_5 = .624; L_M_7 = .412; L_M_8 = .456; L_M_9 = .512</td>
<td>.185</td>
</tr>
<tr>
<td>VC_3</td>
<td>103.485**</td>
<td>33</td>
<td>3.136</td>
<td>.150 (.118 - .182)</td>
<td>.948</td>
<td>.913</td>
<td>L_M_2 = .760; L_M_5 = .946; L_M_6 = .290; L_M_7 = .415; L_M_9 = .643</td>
<td>L_M_2 = .676; L_M_5 = .820; L_M_6 = .311; L_M_7 = .510; L_M_9 = .503</td>
<td>.300</td>
</tr>
<tr>
<td>VC_4</td>
<td>102.839**</td>
<td>34</td>
<td>3.025</td>
<td>.146 (.114 - .179)</td>
<td>.949</td>
<td>.918</td>
<td>L_M_2 = .927; L_M_4 = .305; L_M_5 = .511; L_M_6 = .274</td>
<td>L_M_2 = .762; L_M_4 = .296; L_M_5 = .384; L_M_6 = .255</td>
<td>.281</td>
</tr>
<tr>
<td>VC_5</td>
<td>100.729**</td>
<td>33</td>
<td>3.052</td>
<td>.147 (.115 - .180)</td>
<td>.949</td>
<td>.916</td>
<td>L_M_2 = .652; L_M_3 = .611; L_M_4 = .383; L_M_6 = .433; L_M_8 = .587</td>
<td>L_M_2 = .589; L_M_3 = .585; L_M_4 = .436; L_M_6 = .472; L_M_8 = .577</td>
<td>.196</td>
</tr>
<tr>
<td>VC_6</td>
<td>100.366**</td>
<td>35</td>
<td>2.868</td>
<td>.140 (.109 - .171)</td>
<td>.951</td>
<td>.924</td>
<td>L_M_1 = .546; L_M_5 = .622; L_M_10 = .362</td>
<td>L_M_1 = .487; L_M_5 = .507; L_M_10 = .387</td>
<td>.199</td>
</tr>
<tr>
<td>VC_7</td>
<td>100.445**</td>
<td>35</td>
<td>2.870</td>
<td>.140 (.109 - .173)</td>
<td>.951</td>
<td>.923</td>
<td>L_M_2 = .391; L_M_6 = .298; L_M_8 = .292</td>
<td>L_M_2 = .368; L_M_6 = .339; L_M_8 = .299</td>
<td>.200</td>
</tr>
<tr>
<td>VC_8</td>
<td>108.557**</td>
<td>35</td>
<td>3.102</td>
<td>.149 (.117 - .189)</td>
<td>.946</td>
<td>.914</td>
<td>L_M_2 = .543; L_M_6 = .442; L_M_8 = .367</td>
<td>L_M_2 = .342; L_M_6 = .336; L_M_8 = .251</td>
<td>.200</td>
</tr>
<tr>
<td>VC_9</td>
<td>101.177**</td>
<td>34</td>
<td>2.976</td>
<td>.144 (.112 - .177)</td>
<td>.950</td>
<td>.919</td>
<td>L_M_2 = .543; L_M_3 = .720; L_M_8 = .627; L_M_10 = .473</td>
<td>L_M_2 = .417; L_M_3 = .586; L_M_8 = .523; L_M_10 = .465</td>
<td>.156</td>
</tr>
<tr>
<td>VC_10</td>
<td>110.680**</td>
<td>37</td>
<td>2.991</td>
<td>.145 (.114 - .176)</td>
<td>.945</td>
<td>.918</td>
<td>L_M_6 = .302</td>
<td>L_M_6 = .250</td>
<td>.062</td>
</tr>
</tbody>
</table>

Note: CFI = comparative fit index; d.f. = degrees of freedom; RMSEA = Root mean square error of approximation; TLI = Tucker-Lewis index. L_M_1 = warm; L_M_2 = friendly; L_M_3 = likeable; L_M_4 = similar; L_M_5 = approachable; L_M_6 = knowledgeable; L_M_7 = physique; L_M_8 = advice; L_M_9 = roommate; L_M_10 = friendship. e = error of measurement. *a see Figure 2. ** p < 0.01.
Appendix 1
The Tertiary-Level English Oral Presentation Scale (TEOPS; Adapted from Aryadoust, 2015)

1 = needs improvement 
2 = good 
3 = excellent

<table>
<thead>
<tr>
<th>EVALUATION POINTS</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CONTENT &amp; ORGANISATION</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Interesting opening to capture attention</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Clear context for the controversy leading to thesis/stance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Well-developed presentation of information for explanation or substantiation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Convincing main and supporting ideas; persuasive and sufficient to support stance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Sound analysis in a manner suitable for an audience of educated lay-persons</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Suitable transitional devices to link sections of presentation coherently</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Suitably strong conclusion that advances thesis and points the way forward</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Answered audience’s questions politely, clearly, knowledgeably &amp; effectively</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>VERBAL LANGUAGE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Various sentence patterns &amp; well-formed sentences, correct grammar</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Appropriate vocabulary &amp; style</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Clear &amp; correct pronunciation - stressed syllables &amp; words properly</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Appropriate volume/pitch/tone/speed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>NON VERBAL LANGUAGE</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Enthusiastic and confident– e.g., through words, facial expressions, smiles, voice, tone, etc.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Confident posture – e.g., faced the audience straight &amp; tall - not slouching, not stiff, not pacing every other minute, etc.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Meaningful and fitting gestures – e.g., those that are natural and those that complement the verbal language</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Frequent eye contact with the audience</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Good time management</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Effective audio/visual aids – e.g., relevant PowerPoint slides, and graphs, tables, pictures, objects; were attractively and professionally designed and crafted</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix 2
Likeability Scale (adapted from Reysen, 2005)

Instructions: Please choose how strongly you agree with each statement.

Very Strongly Disagree = 1
Strongly Disagree = 2
Disagree = 3
Neutral = 4
Agree = 5
Strongly agree = 6
Very Strongly Agree = 7

<table>
<thead>
<tr>
<th>This person is.....</th>
<th>Presenter 1</th>
<th>Presenter 2</th>
<th>Presenter 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. warm</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. friendly</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. likeable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. similar to me</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. approachable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. knowledgeable</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. physically attractive</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>I would like this person....</th>
<th>Presenter 1</th>
<th>Presenter 2</th>
<th>Presenter 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>8. for advice</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. as a roommate /co-worker</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

10. I would like to be friends with this person.

Appendix 3
List of the compulsory readings of the module


