Listening. Learning. Leading.

# TOEFL iBT Speaking Test Scores as Indicators of Communicative Language Proficiency

Brent Bridgeman

Donald Powers

Elizabeth Stone

Pamela Mollaun

Educational Testing Service

# What is TOEFL iBT Speaking?

- A measure of speaking ability in typical academic settings
  - Academic course content
  - Campus life
- Emphasis on communicative competence
- Six spoken responses of about one minute each
  - Two independent tasks
  - Four integrated tasks
    - Listen and speak
    - Read and speak
  - Each task rated on 1-4 scale; total score 6-24

# What is SpeechRater™?

- Automated system combining speech recognition tools with natural language processing
- Evaluates extended open-ended speech
  - Other systems evaluate only short, predictable responses
- Currently used to provide scores for TOEFL Practice Online (TPO)

# SpeechRater Components

- Three main components
  - speech recognizer, trained on about 30 hours of non-native speech
  - feature computation module, computing about 40 features predominantly in the fluency dimension
  - scoring model, which combines a selected set of speech features to predict a speaking score using multiple regression
- Assesses primarily fluency and pronunciation

# Relationship of SpeechRater Scores and Expert Human Scores

- Correlation of SpeechRater scores and human scores from expert raters is important component of SpeechRater validity argument
  - In group that took TPO, correlation was 0.57
  - In Field Test data, the correlation was 0.68 (Xi, et al., in press)
- This correlation is only one component of the validity argument

# What is a reasonable criterion for evaluating the validity of ratings from expert humans or a machine?

- TOEFL iBT raters are well trained and highly experienced in evaluating the speech of non-native speakers
  - This is their strength
  - It may also be their weakness
- Undergraduates who are native speakers of English provide an alternative approach to understanding communicative competence

# Two Types of Criterion Scores from Undergraduate Raters

- Comprehension scores and Rating Scales
- Comprehension score
  - Based on responses to multiple-choice questions that require understanding of the spoken response
  - Example:

  Which of the following best expresses the topic of the speaker's response?

  A) An interesting personal experience

  B) An important book

  C) His or her field of study

  D) His or her favorite film

  E) Topic was unclear

# Two Types of Criterion Scores

- Rating Scales—five point ratings
  - Effort
    - "As a listener, how much effort was required to understand the speaker" with choices from "very little" to "not comprehensible"
  - Confidence
    - "How confident are you that you understood what the speaker was trying to say?" with choices from "extremely confident" to "not confident at all
  - Interference
    - How much did the speaker's English language abilities [pronunciation, vocabulary, or grammar] interfere with your understanding of the response? with choices from "Did not interfere at all" to "Almost always interfered"
  - Task Fulfillment

# Example Task Fulfillment Rating

**Now read the question that the speaker was asked:**

**Using the research described by the professor, explain what scientists have learned about the mathematical abilities of babies.**

In your opinion, how successful was the speaker at answering the question?

- Extremely successful
- Very successful
- Somewhat successful
- Not very successful
- Not at all successful

# Undergraduate Raters

- 555 full-time students from:
  - a public university in the Northeast
  - two public universities in the Southeast
  - a private university in the Southeast
  - a private university in the Northwest
- Native speakers of English
- Paid $50 to listen to and evaluate 12 speech samples

# Source of Speech Samples

- Regular examinees from two released operational forms—6 tasks per form
- 94 examinees for each form

# Rating Teams

- Raters assigned to virtual teams—about 5 members to each team
- Each team rated a total of 12 responses--ONE response to each task
  - 6 tasks from Form 1 and 6 from Form 2
- Team score based on the average of the 5 raters for both comprehension items and rating scales
- Score for an examinee based on average across six rating teams—one team for each response

# Mean Comprehension Scores

- For 13 Form 1 questions, Mean = 6.43 (SD = 1.80)
- For 12 Form 2 questions, Mean = 6.01 (SD = 1.68)

# Mean Rating Scale Scores

| Scale | Form 1 | | | | Form 2 | | | |
|---|---|---|---|---|---|---|---|---|
| | # Quest. | M | SD | M/ # Quest. | # Quest. | M | SD | M/ # Quest |
| Effort | 6 | 17.95 | 4.60 | 2.99 | 6 | 17.50 | 4.03 | 2.92 |
| Confidence | 7 | 20.67 | 4.78 | 2.95 | 6 | 17.17 | 3.65 | 2.86 |
| Interference | 6 | 14.14 | 3.93 | 2.36 | 6 | 16.58 | 4.00 | 2.76 |
| Task Fulfillment | 4 | 11.63 | 2.12 | 2.91 | 3 | 8.78 | 1.62 | 2.92 |

# Reliability of Scores (Alpha)

- Operational human scores: 0.87 in both forms
- SpeechRater: 0.95 for Form 1 and 0.96 for Form2
- Comprehension score from undergraduate raters: 0.76 in both forms
- Rating scales

| Rating Scale | Form 1 | Form 2 |
|---|---|---|
| Effort | 0.92 | 0.88 |
| Confidence | 0.91 | 0.85 |
| Interference | 0.92 | 0.89 |
| Task Fulfillment | 0.76 | 0.62 |

# Correlations Among Rating Scales

| Scale | Form 1 | | | Form 2 | | |
|---|---|---|---|---|---|---|
| | Confidence | Interference | Task Fulfillment. | Confidence | Interference | Task Fulfillment |
| Effort | 0.97 | 0.98 | 0.83 | 0.97 | 0.98 | 0.82 |
| Confidence | | 0.95 | 0.80 | | 0.96 | 0.85 |
| Interference | | | 0.80 | | | 0.82 |

# Correlation of Undergraduate Rater Scores with SpeechRater Scores

| | Form 1 | Form 2 |
|---|---|---|
| Undergrad. Rater Scores | SpeechRater | SpeechRater |
| Comprehension | 0.47 | 0.32 |
| Rating Scale Total | 0.45 | 0.28 |

# Correlation of Undergraduate Rater Scores with Operational Human Scores and SpeechRater Scores

| | Form 1 | | Form 2 | |
|---|---|---|---|---|
| Undergrad. Rater Scores | Human | SpeechRater | Human | SpeechRater |
| Comprehension | 0.79 | 0.47 | 0.66 | 0.32 |
| Rating Scale Total | 0.79 | 0.45 | 0.67 | 0.28 |

# Why are correlations higher for Form 1?

| | | Form 1 | | | Form 2 | | |
|---|---|---|---|---|---|---|---|
| Language | Score | *n* | Operational | SpeechRater | *n* | Operational | SpeechRater |
| Asian[a] | Comprehension | 34 | 0.71 | 0.31 | 48 | 0.60 | 0.25 |
| | Rating Total | | 0.73 | 0.28 | | 0.76 | 0.32 |
| European[b] | Comprehension | 33 | 0.80 | 0.52 | 7 | 0.91 | 0.81 |
| | Rating Total | | 0.88 | 0.62 | | 0.93 | 0.74 |
| Other | Comprehension | 27 | 0.66 | 0.31 | 39 | 0.66 | 0.26 |
| | Rating Total | | 0.52 | 0.14 | | 0.52 | 0.11 |

[a]Chinese, Japanese, and Korean. [b]French, German, and Italian

# Crosstab of Examinees with High and Low Comprehension Scores By High and SpeechRater Scores on Form 1

| | Comprehension--High | Comprehension—Low |
|---|---|---|
| SpeechRater--High | 18 | **5** |
| SpeechRater--Low | **3** | 19 |

Note.—“High” is top third and “Low” is bottom third; middle third is omitted.

# Crosstab of Examinees with High and Low Comprehension Scores By High and Low Human and SpeechRater Scores on Form 1

| | Comprehension--High | Comprehension—Low |
|---|---|---|
| Human--High | 20 | **0** |
| Human--Low | **0** | 25 |
| SpeechRater--High | 18 | **5** |
| SpeechRater--Low | **3** | 19 |

Note.—"High" is top third and "Low" is bottom third; middle third is omitted.

# Crosstab of Examinees with High and Low Rating Total Scores By High and Low SpeechRater Scores on Form 1

| | Rating Scale Total--High | Rating Scale Total--Low |
|---|---|---|
| SpeechRater--High | 18 | **7** |
| SpeechRater--Low | **4** | 18 |

Note.—"High" is top third and "Low" is bottom third; middle third is omitted.

# Crosstab of Examinees with High and Low Rating Total Scores By High and Low Human and SpeechRater Scores on Form 1

| | Rating Scale Total--High | Rating Scale Total--Low |
|---|---|---|
| Human--High | 20 | **1** |
| Human--Low | **0** | 26 |
| SpeechRater--High | 18 | **7** |
| SpeechRater--Low | **4** | 18 |

Note.—"High" is top third and "Low" is bottom third; middle third is omitted.

# Crosstab of Examinees with High and Low Comprehension Scores By High and Low SpeechRater Scores on Form 2

| | Comprehension--High | Comprehension--Low |
|---|---|---|
| SpeechRater--High | 11 | **8** |
| SpeechRater--Low | **9** | 13 |

Note.—"High" is top third and "Low" is bottom third; middle third is omitted.

# Crosstab of Examinees with High and Low Comprehension Scores By High and Low Human and SpeechRater Scores on Form 2

| | Comprehension--High | Comprehension--Low |
|---|---|---|
| Human--High | 19 | **2** |
| Human--Low | **6** | 20 |
| SpeechRater--High | 11 | **8** |
| SpeechRater--Low | **9** | 13 |

Note.—"High" is top third and "Low" is bottom third; middle third is omitted.

# Conclusions

- Study provides strong evidence supporting the validity of TOEFL iBT scores as a measure of communicative competence.
  - Undergraduate students without any training in evaluating speech of international students rated their comprehension of speech samples in a manner that was highly consistent with the way examinees were ordered by experienced TOEFL iBT raters

# Conclusions

- SpeechRater is reliably assessing some portion of the speaking construct
  - but is not assessing the full construct in a way that is consistent with human evaluations of speaking competence, whether those human judgments are from experienced professionals or naïve undergraduates