ISSN 1930-9317

TOEFL iBT™ Research Report

TOEFLiBT-11
August 2009

How Do Raters From India Perform in Scoring the TOEFL iBT™ Speaking Section and What Kind of Training Helps?

Xiaoming Xi

Pam Mollaun

How Do Raters From India Perform in Scoring the TOEFL iBT™ Speaking Section and What Kind of Training Helps?

Xiaoming Xi and Pam Mollaun
ETS, Princeton, New Jersey

RR-09-31

Abstract

This study investigated the scoring of the Test of English as a Foreign Language™ Internet-based Test (TOEFL iBT™) Speaking section by bilingual or multilingual speakers of English and 1 or more Indian languages. We explored the extent to which raters from India, after being trained and certified, were able to score the Speaking section for TOEFL iBT examinees with mixed first language (L1) backgrounds, especially those speaking an Indian language, accurately and consistently. The effectiveness of a special training package designed for scoring Indian examinees was examined as well. A total of 26 trained and certified raters from India were randomly divided into 2 groups and participated in 2 on-site scoring sessions in Mumbai. In the first session, both groups received regular training for scoring the TOEFL iBT Speaking section, which was largely similar to that received by raters in North America. In the second scoring session, 1 group continued to receive the regular training while the second group was trained using a special training package. Rater feedback surveys were also given to the raters. It was found that with training similar to that which operational U.S.-based raters receive, the raters from India performed as well as the operational raters in scoring both Indian and non-Indian examinees. In addition, the special training helped the raters score Indian examinees more consistently, leading to increased score reliability estimates. It also boosted raters' levels of confidence in scoring Indian examinees.

Key words: TOEFL iBT, speaking, speech scoring, rater background characteristics, non-native raters, rater bias

The Test of English as a Foreign Language™ (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board® assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations® (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education. The test is now wholly owned and operated by ETS.

ETS administers the TOEFL program under the general direction of a policy board that was established by, and is affiliated with, the sponsoring organizations. Members of the TOEFL Board (previously the Policy Council) represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, two-year colleges, and nonprofit educational exchange agencies.

❖ ❖ ❖

Since its inception in 1963, the TOEFL has evolved from a paper-based test to a computer-based test and, in 2005, to an Internet-based test, TOEFL iBT. One constant throughout this evolution has been a continuing program of research related to the TOEFL test. From 1977 to 2005, nearly 100 research and technical reports on the early versions of TOEFL were published. In 1997, a monograph series that laid the groundwork for the development of TOEFL iBT was launched. With the release of TOEFL iBT, a TOEFL iBT report series has been introduced.

Currently this research is carried out in consultation with the TOEFL Committee of Examiners. Its members include representatives of the TOEFL Board and distinguished English as a second language specialists from the academic community. The Committee advises the TOEFL program about research needs and, through the research subcommittee, solicits, reviews, and approves proposals for funding and reports for publication. Members of the Committee of Examiners serve four-year terms at the invitation of the Board; the chair of the committee serves on the Board.

Current (2008-2009) members of the TOEFL Committee of Examiners are:

Alister Cumming (Chair)	University of Toronto
Geoffrey Brindley	Macquarie University
Frances A. Butler	Language Testing Consultant
Carol A. Chapelle	Iowa State University
John Hedgcock	Monterey Institute of International Studies
Barbara Hoekje	Drexel University
John M. Norris	University of Hawaii at Manoa
Pauline Rea-Dickins	University of Bristol
Steve Ross	Kwansei Gakuin University
Mikyuki Sasaki	Nagoya Gakuin University
Robert Schoonen	University of Amsterdam
Steven Shaw	University of Buffalo

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

Table of Contents

Appendixes

List of Tables

Executive Summary

Although nonnative raters are frequently used in many large-scale speaking assessments, there has been inconclusive evidence regarding the impact of raters' familiarity with examinees' L1 on their evaluations of examinees' speaking proficiency. Some research has shown that familiarity with the speaker's accent facilitates comprehension and may thus lead to more lenient evaluations of the overall speech quality (Brodkey, 1972; Smith & Bisazza, 1982; Smith & Rafiqzad, 1979), whereas other studies have revealed lower tolerance of their peers' speech by nonnative speakers of the target second language (L2) than native speakers (Fayer & Krasinski, 1987; Sheorey, 1985). These conflicting findings may be partially due to differences in whether naive or trained raters are used, how adequately the raters are trained, and whether the raters undergo rigorous certification requirements. In particular, few investigations have looked into whether rigorous training and certification procedures could minimize potential bias introduced by greater exposure to the language of examinees with a particular L1. Also, no previous research has examined what kind of training may mitigate the potential negative effects of raters' familiarity with examinees' L1.

This study investigated the scoring of the Test of English as a Foreign Language™ Internet-based test (TOEFL iBT™) Speaking section by bilingual or multilingual speakers of English and one or more Indian languages. It attempted to explore whether raters from India, after being trained and certified, were able to score TOEFL iBT examinees with mixed L1 backgrounds, especially those speaking an Indian language, accurately and consistently despite their greater familiarity with the Indian accents than other raters. The effectiveness of a special training package designed for scoring Indian examinees was examined as well.

The raters from India were tested for their speaking proficiency, went through an extensive online training program, and each group completed a rater certification test. The 26 selected raters were randomly divided into two groups and participated in two on-site scoring sessions in Mumbai. In the first session, both groups received identical training for scoring the TOEFL iBT Speaking section. This generic training was largely similar to that received by U.S.-based raters and included review and practice scoring of responses from mixed L1 speakers to three generic tasks. After being trained on the generic tasks, raters rated 100 responses of speakers of Indian and non-Indian languages on each of three similar tasks selected for this study. In the second scoring session, the first group continued to receive generic training while

the second group received additional training using a special training package. This special training involved using a set of benchmarks and calibration samples of Indian examinees.

The correlations and weighted kappas between the scores assigned by the raters from India and those previously assigned by operational raters were computed for different groups across the two scoring sessions. In addition, generalizability (G) studies were conducted to examine the reliability of the scores assigned to Indian and non-Indian responses for the two rater groups across the two scoring sessions. A rater feedback survey was given to all of the raters at the end of the first scoring session to elicit their feedback on the training and their scoring experiences. Another survey, focusing on the effectiveness of the special training, was given to the second group after their completion of the second scoring session.

It was found that with training similar to that which operational raters receive, the raters from India performed as well as the operational raters in scoring both Indian and non-Indian examinees. This is evident in the high agreement between the scores assigned by the raters from India and the operational raters. The special training did not give the intervention group an advantage when their scores were compared with the operational raters' scores at the task level; however, when the scores were summed across the three tasks, the special training group had slightly better agreement with the operational raters than did the group that received generic training. In addition, the special training helped the raters score Indian examinees more consistently, leading to increased score reliability estimates. It also boosted raters' levels of confidence in scoring Indian examinees.

The results suggest that it is appropriate to use raters from India who have similar qualifications as those in this study for scoring TOEFL iBT Speaking. In addition, use of Indian benchmarks and calibration samples along with mixed-L1 ones is recommended for training raters from India for scoring TOEFL iBT Speaking.

Introduction

The Test of English as a Foreign Language™ Internet-based test (TOEFL iBT™) test made its debut in September 2005 in North America and was later launched in other countries and regions. With the test volumes going up, it has become necessary for the TOEFL® program to expand the rater pool for the Speaking section to ensure efficiency in scoring and score reporting.

Traditionally, large-scale speaking assessments at ETS, including the Test of Spoken English™ (TSE®) and the Test of English for International Communication™ (TOEIC®) Speaking test, have used primarily native English-speaking raters in America. However, we also realize that some English language training professionals in TOEFL test candidates' home countries may be qualified raters. The TOEFL iBT test has been designed to support the teaching and learning of academic English worldwide, and it seems consistent with this goal to engage English language teaching and assessment specialists in candidates' home countries in scoring the Speaking section. By including them in the assessment process, the TOEFL program could promote a better understanding of the test content and scoring criteria, which is expected to positively impact classroom teaching practices.

Another perspective comes from the current debate about whether educated speakers of standard varieties of English (e.g., British English, American English) should set the norms for English teaching and testing (Davies, Hamp-Lyons, & Kemp, 2003; Kachru, 1986; Quirk, 1985, 1990). Advocates of the World Englishes (WEs) view would support using English-speaking raters with diverse L1 backgrounds for scoring English language assessments, based on the argument that standard English norms should not be used to evaluate the English proficiency of learners. However, since the TOEFL iBT test measures English language abilities required to handle academic studies in English-medium universities (e.g., in America, Canada, the United Kingdom, New Zealand, Australia), it adopts standard English norms by educated speakers in the evaluation of the writing or speaking sections (although the highest performance levels described in the writing or speaking scoring rubrics emphasize the overall effectiveness of the written or spoken performance rather than native-like performance).

Despite the potential advantage of using English teaching and assessment specialists in candidates' home countries, critical issues remain to be examined, as these specialists may include both native and highly proficient nonnative speakers of English who are familiar with certain accents. Therefore, it is important to understand the impact of raters' familiarity with the candidates' L1 and other rater background characteristics on the quality of their scoring. Another issue is whether highly proficient nonnative speakers of English may score examinees of mixed L1 backgrounds differently than native-speaking raters. To investigate these issues, we conducted two studies to investigate the scoring of TOEFL iBT Speaking by raters in China and India, respectively. This report focuses on the India study.

Although nonnative raters are frequently used in large-scale speaking assessments, there have been inconsistent findings regarding the impact of raters' familiarity with examinees' L1 on their evaluations of examinees' speaking proficiency. Some research has shown that familiarity with the speaker's accent may lead to more lenient evaluations of the overall speech quality (Brodkey, 1972; Gass & Varonis, 1984; Smith & Bisazza, 1982; Smith & Rafiqzad, 1979). However, other studies have demonstrated lower tolerance of their peers' speech by nonnative speakers of the target L2 than native speakers (Fayer & Krasinski, 1987; Sheorey, 1985). To further complicate the situation, some research has supported the finding that naive listeners do not show a clear and consistent advantage in understanding speech produced in their own accent (Major, Fitzmaurice, Bunta, & Balasubramanian, 2002; Munro, Derwing, & Morton, 2006). These conflicting findings may be partially due to differences in whether naive or trained raters are used, how adequately the raters are trained, and whether the raters are certified. In particular, few investigations have looked into whether rigorous training and certification procedures could minimize potential bias introduced by greater exposure to examinees' L1. Also, no previous research has examined what kind of training may mitigate the potential negative effects of raters' familiarity with examinees' L1.

This study investigates the scoring of the TOEFL iBT Speaking section by bilingual or multilingual speakers of English and one or more Indian languages. It attempts to explore whether trained and certified raters from India are able to score TOEFL examinees with mixed L1 backgrounds, especially Indian examinees, accurately and consistently. The effectiveness of a special training package designed for scoring Indian examinees is examined as well.

Impact of Rater Background Characteristics on the Scoring of Speaking

Previous research on the assessment of second or foreign language learners' speaking proficiency by evaluators with different backgrounds has looked at untrained native versus nonnative evaluators (Brodkey, 1972; Caban, 2003; Fayer & Krasinski, 1987; Gorosch, 1973;

Kim, 2009; Sheorey, 1985; Smith & Bisazza, 1982; Smith & Rafiqzad, 1979), trained native versus nonnative raters (Brown, 1995), laypeople versus professionals with training in a second or foreign language (Barnwell, 1989; Caban, 2003; Chalhoub-Deville, 1995; Galloway,1980; Gorosch, 1973; Hadden, 1991), male versus female raters (Eckes, 2005), raters with linguistic versus occupational backgrounds (Brown, 1995), and raters who are native speakers of different standard varieties of English (Chalhoub-Deville & Wigglesworth, 2005).

The investigations of native versus nonnative perceptions of nonnative communication are the most relevant to the present study. However, inconclusive evidence has been established regarding the impact of evaluators' native language backgrounds on their evaluation of nonnative speech, as discussed above. Also, the bulk of the research involves explorations of naive listeners' perceptions of nonnative speech in a nonassessment context. The focus is to understand how naive listeners who share a similar linguistic background as the speaker or have a different linguistic background perceive various properties of nonnative speech produced by the speaker. However, in explorations of trained rater perceptions in an assessment context, the paramount goal is to reduce potential listener or rater bias and ensure fairness through rater training. In this section, we only review studies that use fairly clear scoring rubrics and raters with training in a second or foreign language or trained raters, which are typically required in an assessment or classroom evaluation context.

Galloway (1980) investigated the perceptions of the communicative efforts of 10 university-level American students of Spanish based on their responses to a general knowledge question by four groups of evaluators. These four groups were made up of native high school Spanish teachers, nonnative high school Spanish teachers, nonteaching native speakers living in the United States with a fair-to-good command of English, and nonteaching native speakers living in Spain with no or poor command of English. The students were rated on five subcategories: amount of communication, efforts to communicate, comprehensibility, paralanguage, and overall impression. Overall, no significant differences were found across the four groups on the five subscales. Regarding specific aspects of speech, the nonteaching native speakers of Spanish living in the United States were less disturbed by pronunciation than the other groups. The nonnative high school Spanish teachers seemed to be more bothered than the nonteaching native speakers of Spanish by the slowness with which the students spoke. In addition, their written comments about each student's performance revealed that nonnative teachers focused primarily on grammatical accuracy, whereas the nonteaching native speakers were drawn more to the content or message. Although this study has offered some interesting exploratory results for future research to follow up on, the small sample size (10 students) has seriously constrained the generalizability of the findings. The scales used in this study are also far less elaborated and explicit than typical rating scales in a more formal assessment context, so the results may have limited applicability to other contexts.

Caban (2003) conducted a study where a total of 83 *untrained* Japanese L1 or English L1 raters with or without English as a second language/English as a foreign language (ESL/EFL) background rated four Japanese students' English oral interviews on seven categories (fluency, grammar, pronunciation, comprehension techniques, content of utterance, language appropriateness, and overall intelligibility). It was found that the English L1 speaker groups were consistently more lenient in evaluating the pronunciation quality of Japanese-accented English than Japanese L1 raters. Another finding was that the ESL/EFL-trained Japanese L1 raters rated pronunciation and grammar more harshly but compensation techniques, language appropriateness, and overall intelligibility more leniently than the other groups. The Japanese L1 speakers also exercised more leniency in scoring fluency and grammar than the English L1 raters. However, a few serious limitations rendered the findings of the study questionable. First, despite the effort to use a large number of raters, this study used an extremely small scoring sample (4 students' oral interviews), which called into question the suitability of the many-facet Rasch model (MFRM) procedures used in the study. In addition, the relatively nonvarying proficiency levels represented by these four Japanese students seriously limits the generalizability of the findings. The rating scale, with as many as seven categories and 15 points to differentiate among, would daunt any trained raters, let alone untrained naive raters.

Brown (1995) looked at the scoring performance of 33 native and nonnative raters of a Japanese test for tourist guides taken by examinees who speak Japanese as a second or foreign language. The raters, who were provided with clear assessment criteria and were adequately trained, scored 51 examinees. Using MFRM as the primary analytic tool, Brown found that although the overall rater severity did not differ significantly across these two rater groups, nonnative raters were harsher in evaluating some aspects of speech, including politeness and pronunciation. That is to say, although the overall assessment of candidates' proficiency would not change depending on which rater group was used, the two rater groups were probably going through different thought processes to arrive at similar scores. She speculated that the nonnative raters may have held the examinees to higher standards on politeness than the native raters did due to the arduous learning process they themselves had gone through. She attributed the discrepancies in assessing pronunciation to the different foci of the two rater groups and argued that native speakers only penalized candidates for errors that seriously hindered communication, while nonnative speakers marked down candidates for any pronunciation errors. Since these interpretations were speculative, she called for qualitative research to cast light on the interpretations of the findings.

Kim (2009) is one of the few studies to investigate how trained native versus nonnative English–speaking teachers perform in scoring an English speaking test. In her study, 12 native-speaking English teachers in Canada and 12 Korean-L1 English teachers with graduate degrees in linguistics or language education scored a total of 80 spoken responses from 10 Korean speakers to a computerized oral test, after being trained with sample responses. Using MFRM, Kim did not find any total test score differences between these two groups. In addition, neither group showed any positive or negative bias[1] toward a particular task or task type. However, based on raters' written justifications of scores, she found that native and nonnative teachers may assign similar scores to the same responses but for somewhat different reasons. The author speculated that more rigorous rater training may help reduce the difference in the decision-making processes across the two groups.

In all of these studies reviewed above, although the number of raters used is generally adequate, the numbers of examinees scored were very small, ranging from 4 to 51. These small samples limit the generalizability of the results and also call into question the appropriateness of using the MFRM procedures for some of the studies.

While Brown (1995) and Kim (2009) are most pertinent to the current study because they used clearly defined scoring rubrics and trained raters, these studies were conducted in different test contexts and employed raters with different L1 backgrounds. Raters' perceptions and orientations may well change depending on the features of the speaking tasks, the emphases in the scoring rubrics, the nature of candidates' speech, and the rigor of rater training. In light of the paucity of speech scoring research that combines quantitative and qualitative methodologies, the present study also employed qualitative techniques to inform the interpretation of the quantitative results. Additionally, raters from India generally speak a variety of the English language commonly referred to as *Indian English* in addition to one or more Indian languages, and are thus quite different than the raters investigated in previous research with respect to the impact of L1 familiarity on speech scoring. As is known, Indian English refers to the dialects or varieties of English spoken widely in India. It has evolved into an English variety of a unique flavor (e.g., peculiarities in its phonological and prosodic patterns, syntax, and vocabulary) due to the blending of British English and Indian languages and dialects. However, British English is an official language of central and state governments in India; Indian English, although widely spoken, is not considered proper usage by either government-related institutions (such as offices and schools) or educated Indians (Indopedia, 2004).

Effect of Rater Training on Scoring Performance

Many large-scale writing or speaking assessments implement procedures to train, certify, and monitor their raters to ensure scoring quality. Sometimes alternative training procedures are investigated to either improve the efficiency of training or to increase the accuracy and consistency of scoring. Research on the effects of rater training programs has focused on the scoring of writing tests. While some studies compare the impact of rater training using naive and trained raters, others are conducted for the purpose of modifying existing training procedures that target regular raters. Both types of studies are reviewed.

Shohamy, Gordon, and Kraemer (1992) investigated the impact of rater training on the scoring of a writing test. A total of 20 raters participated in this study; half were English teachers (professionals) and the other half did not have any training in English teaching (laypersons). Half of the professionals and laypersons received training and the other half did not. Then all of the raters rated 50 writing samples. Shohamy and her associates found that the scores given by the trained raters were more reliable than those by the untrained, as indicated by the overall inter-rater reliability estimates. A repeated measures analysis of variance (ANOVA) also showed that training had significant effects on raters’ scores.

To explore the effects of training, Weigle (1998) used raters both with and without experience to score a writing test used for placement purposes at a major American university. Two weeks before the training session, all of the raters rated different but overlapping sets of 15 essays, each from two writing tasks. Then they participated in a training session of 90 minutes and rated compositions for 6 to 10 hours as part of the operational scoring over a period of 10 days (this scoring data was not analyzed in her study). Following the operational scoring, they were asked to score again different but overlapping sets of 16 essays on each of the two writing prompts within 1 to 3 weeks, most of which had not been scored by the raters prior to the training session. Weigle used MFRM to analyze the pre- and post-training scoring data and found that the raters, especially the new raters, were less varied in their overall severity levels after the training. Another positive effect was that rater consistency improved from pre- to post-training scoring sessions. Despite the positive effects, the training session did not completely eliminate variation in rater severity levels. The generalizability of the findings was constrained by a few limitations. The small sample used (16 essays on each task) was certainly a limitation. Another caveat was that all the raters had participated in about 6 to 10 hours of operational scoring before scoring for this study. Familiarity developed during the operational scoring could have confounded any training effects.

Elder, Barkhuizen, Knoch, and von Randow (2007) examined the effects of an online rater training program on raters' scoring of a writing test. This study was part of a larger study consisting of several scoring sessions that investigated both the effectiveness of the online training program and individualized rater feedback. Elder et al. (2007) reported on the effectiveness of the online training program and Elder, Knoch, Barkhuizen, and von Randow (2005) focused on the effects of individual rater feedback. In Elder et al. (2007), eight accredited raters participated in this study. The regular face-to-face training involved interactive discussions of a set of benchmark scripts, as well as independent scoring of them and comparison of one's own scores to the official scores. In the online training program, the raters scored a few prescored benchmark scripts, and for each script they received feedback on the difference between the official score and their score, and commented in writing why they thought there was a discrepancy. They were also encouraged to review the rationale that explained why a benchmark script was assigned a certain score. They then rated 10 more benchmark scripts and were encouraged to keep scoring until they felt confident about the scores they assigned. These eight raters rated a randomly selected sample of 100 writing scripts on four different writing tasks over a period of a week; then they went through the online training program described above and completed a survey that elicited their reactions to the training; and, finally, they re-rated 50 of the 100 original scripts scrambled in order within a week. An MFRM analysis of their before- and after-training scoring data showed little gain in the overall reliability of their scores as a result of the online training. Raters' reactions to the program were also mixed, some being positive with others offering suggestions for improving the program or expressing preference for the face-to-face training.

Elder et al. (2005) examined the effects of individual rater performance feedback on the scoring of the same writing test employed in Elder et al. (2007). The eight raters began by scoring 100 Diagnostic English Language Needs Assessment (DELNA) scripts, received online training, and then scored 50 randomly sampled DELNA writing scripts. Individualized feedback for each rater was produced based on an MFRM analysis of their scores, which included information on the rater's biases associated with the scoring of a dimension of the analytic criteria, written feedback analyzing the rater's overall scoring performance in the first two rounds of scoring, and his or her severity in comparison to other raters. This individualized feedback was provided to each rater, following general explanations to them as a group about how it should be interpreted. Then they scored 62 to 64 scripts from the 100 original scripts. Similar rater severity and bias statistics were obtained on this batch of scoring data using the MFRM analysis and compared to those generated based on the prior-feedback scoring data. The results showed that the individual rater performance feedback led to less variation in raters' severity levels. The bias displayed in scoring the content or fluency dimension by four out of the six raters disappeared after this special training with customized rater performance feedback. The postfeedback rater questionnaire responses also revealed that the raters were very positive about the opportunity to review feedback tailor-made to their scoring patterns and perceived it as a useful technique to improve their awareness of their own scoring behavior.

Knoch, Read, and von Randow (2007) conducted a follow-up study on the same writing assessment to compare the effects of online and face-to-face training. Sixteen raters rated a set of 70 writing samples in the first phase, and were then split into the online and face-to-face groups and received different types of training. The online training program had a few improvements compared to the one used in Elder et al. (2007), but the training procedures remained largely the same as reported in Elder et al. (2007). The face-to-face group received customized feedback based on an MFRM analysis of their scoring performance during phase 1 and also had a chance to ask a researcher questions about the feedback. They then rated 15 scripts on their own at home and discussed each script in the face-to-face training session. Finally, both groups re-rated the same 70 scripts from phase 1 presented in a different order. Using MFRM analyses, they concluded that both groups showed reasonable scoring consistency before and after training, but that the online group demonstrated more consistency after training than the face-to-face group. Furthermore, neither group showed a halo effect after training, but the face-to-face group was more successful in moving away from assigning similar scores to different scripts. In terms of individual biases, raters in both groups were able to reduce some previous biases identified in the pretraining scoring phase, but new biases occurred as well. The face-to-face groups seemed to have particularly benefited from the individualized feedback and were able to reduce all previous biases. The overall conclusion with these results taken together was that the two training programs were equally effective and neither method showed a clear advantage over the other. This study suffered from the limitation that the raters were not strictly randomly assigned to the two groups.

Three of the five studies reviewed above explored the effects of a conventional training method or procedure, with the exception of Elder et al. (2005) and Knoch et al. (2007), which looked into the use of customized rater feedback. Weigle (1998) used a pre–post-training design with a single rater group, and Shohamy et al. (1992) employed a study design that involved an experimental group and a control group. The other two studies by Elder and her associates examined the effectiveness of alternative training methods but did not use a control group. Thus it was not possible to tease out the effects of the training from the effects of more practice in scoring. In Knoch et al., the focus was comparing online and face-to-face training programs, but because customized rater feedback was included as part of the face-to-face training, it was not possible to isolate the effects of the individualized feedback and the on-site interactive training. The present study investigates the effectiveness of an unconventional training method. A control group was also used to disambiguate potential alternative interpretations of the findings.

Although MFRM has been used in most of the studies reported above, this study employs G theory (Brennan, 2001; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) as the major analytic technique. G theory is a well-established methodology commonly used to estimate score reliability and inform decisions about measurement designs (e.g., how many tasks and how many ratings per task should be obtained from raters randomly drawn from a pool to achieve a desirable level of score reliability). Since the emphasis of this study is to investigate the overall scoring behavior and performance of a rater group under certain training conditions, rather than the scoring performance of individual raters, G theory is better suited as the analytic technique.

Research Questions

This study attempts to determine whether adequate training and certification procedures could reduce potential scoring inconsistencies associated with raters' greater exposure to examinees' L1s. Questionnaires were used to elicit raters' feedback on their training and scoring experiences. The research questions addressed are as follows:

1. How did the raters from India perform in scoring the TOEFL iBT Speaking section with training similar to that received by operational raters?
2. To what extent did the special training impact the quality of scores assigned by the raters from India?
3. How did the raters from India perceive their scoring experiences at the end of Scoring Session 1?
4. How did the raters from India perceive the effectiveness of the special training?

Method

Speaking Section of the TOEFL iBT Test

The Speaking section of the TOEFL iBT test measures examinees' English oral communication skills for studying in English-medium colleges and universities. It consists of six speaking tasks (Table 1). The first two tasks are independent tasks that ask the examinees to speak about familiar topics. The remaining four are integrated tasks that require examinees to listen and speak or read, listen, and speak. Two of them involve a campus-based situation, and the other two involve an academic topic. The listening and reading materials are short and memorable. The test is approximately 20 minutes long. For each of the six questions, examinees are given 15 to 30 seconds to prepare a response. Response time allowed for each question ranges from 45 to 60 seconds.

The Holistic Scoring Rubrics

The scoring rubric for TOEFL iBT Speaking contains descriptors for three dimensions: Delivery, Language Use, and Topic Development (see Xi & Mollaun, 2006, for the scoring rubric). The raters issue a *holistic* score for each response on a scale of 0-4 that is based on these three dimensions. Delivery refers to the pace and clarity of the speech. In assessing delivery, raters consider the speakers' pronunciation, intonation, rate of speech, and degree of hesitancy.

Language Use refers to the range, complexity, precision, and automaticity of vocabulary and grammar use. Raters evaluate candidates' ability to select words and phrases and to produce structures that appropriately and effectively communicate their ideas. Topic Development refers to the coherence and fullness of the response. When assessing this dimension, raters take into account the progression of ideas, the degree of elaboration, the completeness, and, in integrated tasks, the accuracy of the content. As specified in the holistic rubric, raters are instructed to follow these guidelines during holistic scoring: An examinee has to be on target for all three dimensions to receive a score of 4, and for at least two of the dimensions to receive a score of 1, 2, or 3. These guidelines helped raters make overall holistic judgments.

Table 1

The Six Tasks in the Speaking Section of the TOEFL iBT Test

	Task type	Topic	Planning time (in seconds)	Response time (in seconds)
Task 1	Independent	Familiar topics	15	45
Task 2	Independent	Familiar topics	15	45
Task 3	Integrated (reading-listening-speaking)	Campus life	30	60
Task 4	Integrated (reading-listening-speaking)	Academic course content	30	60
Task 5	Integrated (listening-speaking)	Campus life	20	60
Task 6	Integrated (listening-speaking)	Academic course content	20	60

The Raters

The raters in this study were selected through a multistep process similar to that used to select operational TOEFL iBT Speaking raters, illustrated in Figure 1. In the first step, 53 bilingual or multilingual speakers of English and one or more Indian language or dialect with some English teaching and/or speaking assessment scoring experience were invited to complete a Web-based background questionnaire. Only participants who had a master's or doctoral degree with experience teaching English to Indian students were selected to continue onto the next stage. In the second step, the selected participants completed a TOEFL iBT Speaking test. Those who scored 23 or higher (on the 0-30 scale) were then instructed to go through the online rater training tutorial and take a rater certification test, both used for screening operational TOEFL iBT raters. Like the operational raters, the participants had two opportunities to pass the test: if they failed the first one, they were instructed to review the training tutorial again before taking the second test. Each certification test includes 30 candidate responses, representing a range of score levels and L1s on three representative TOEFL iBT Speaking tasks. To pass the certification test with the standard criteria for operational raters, they had to agree perfectly with the official scores on 70% of the responses and must have no discrepant scores[2] (i.e., scores that differed from the official scores by 2 points or more). To obtain a larger rater sample, we decided to relax the passing criteria by allowing those with 50% perfect agreement and no discrepant scores to participate in the study. We refer to this criterion as the lenient criterion.

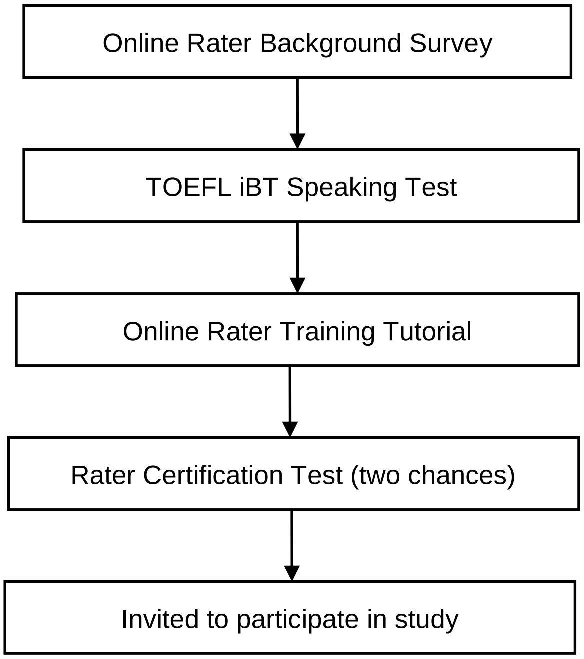
Online Rater Background Survey
TOEFL iBT Speaking Test
Online Rater Training Tutorial
Rater Certification Test (two chances)
Invited to participate in study

***Figure 1*: Rater selection process.**

Following the procedure described above, 26 Indian speakers were invited (some of the qualified raters were not able to participate due to time conflicts) to participate in the final on-site training and scoring activities that occurred in Mumbai. All of them reported having English teaching experience at various levels (high school, college, or commercial language schools).

Twenty-one of them were English teachers or trainers at the time of the study; others were examiners, consultants, or freelance writers.

The raters provided information on their native language, and familiarity with other Indian languages, in a background questionnaire. The results show nine different native languages were represented by the 26 raters: Gujarati (7), Tamil (5), Hindi (3), Kannada (3), Marathi (3), Malayalam (2), Bengali (1), English (1), and Sindhi (1). As will be discussed later in the section on the scoring samples, the Indian examinees scored by the participating raters are primarily Hindi (a major Northern Indian language) and Tamil or Telugu (major Southern Indian languages) speakers.

When asked about the major native languages of students they have most often worked with or people they have had the most contact with (the top three), all of the raters reported Hindi, and 10 reported Tamil (although none reported Telugu). In addition, 23 of them reported being at basic, proficient, or advanced levels in listening and speaking in Hindi, and 7 in Tamil (they were required to self-assess their levels in three languages only). Although none included Telugu in their responses to the questions above, when asked how difficult it is for them to understand the spoken English of heavily accented speakers with Southern Indian accents (1 indicating *very easy* and 4 *very difficult*), 18 of them gave it a 1 and only 8 of them gave it a 2. This indicates that they perceived little difficulty in understanding the accents of Southern Indian language speakers. When it comes to Northern accents, 22 provided a rating of 1 and 4 rated it at 2, suggesting greater ease in understanding the accents of Northern Indian language speakers.

The raters’ native language background as reported above represents a variety of languages spoken in India. The survey responses related to familiarity with and exposure to other Indian languages and knowledge of languages other than their native language also show that the raters selected for this study came from multilingual language backgrounds. Furthermore, the raters were based in Mumbai—a large, ethnically diverse city—and thus may have had more opportunities to encounter a variety of different Indian languages and accents.

Taken together, it is safe to assume that the raters in this study are reasonably familiar with a variety of major Indian languages.

As for their speaking proficiency, all of them scored 23 or higher on TOEFL iBT Speaking practice test and 21 scored perfect or nearly perfect (28 or higher). Fifteen of them passed the certification test with the standard criteria and 11 with the lenient criteria (hereafter referred to as pass and near-pass raters, respectively).

All of the participating raters were compensated for their participation. They were made aware of the general purpose of the study. However, we made it clear that if ETS decided to use raters from India for scoring the TOEFL iBT Speaking section, they would need to be retrained and recertified. Therefore, satisfactory performance in this research study would not guarantee future employment with ETS as a TOEFL iBT Speaking rater.

General Procedure of the Study

The general procedure of the study is shown in Table 2. The 26 selected raters were divided into two groups of 13, with each group roughly equally matched in certification results (pass and near-pass). Each group participated in two on-site scoring sessions in Mumbai. Each scoring session spanned 2 days. In the first session (Day 1 and Day 2), both groups received regular training, which was largely similar to that received by operational raters. After being trained on each of three items, they rated 100 responses in English of speakers of Indian and non-Indian languages (a total of 300 responses on three items). At the end of Scoring Session 1, they filled out a rater feedback survey. In Scoring Session 2 (Day 3 and Day 4), the first group continued to receive the regular training (regular training group) while the second group was trained using a special training package (special training group). After Scoring Session 2 was completed, raters in the special training group were instructed to fill out a second rater feedback survey. The scoring samples, training procedures, and survey instruments used are described in detail below.

Table 2

General Procedure of the Study

	Rater Group 1 (Regular training group)	Rater Group 2 (Special training group)
Scoring Session 1	Regular training and scoring Rater Feedback Survey 1	
Scoring Session 2	Regular training and scoring	Special training and scoring Rater Feedback Survey 2

The Scoring Samples

Two scoring samples taken from the operational TOEFL iBT test data were selected and used in Scoring Sessions 1 and 2, respectively. Only responses scored in the range of 1-4 were included in this study. The two scoring samples were similar in score distribution (see Appendices A and B) and L1 representation, but populated by different examinees.

Each scoring sample consisted of 100 examinees' responses to three TOEFL Speaking tasks (Tasks 2, 4, and 5), with 300 responses in total. In Scoring Session 1, 50 examinees were Indian (21 Hindi, 16 Telugu, 12 Tamil, and 1 Punjabi) and most of the rest were evenly distributed amongst major languages representative of the TOEFL population (9 Arabic, 10 Chinese, 8 French, 10 Korean, and 10 Spanish). In Scoring Session 2, the frequency of examinees by native language was similar to that of Session 1: 52 examinees were Indian (24 Hindi, 18 Telugu, and 10 Tamil), 9 Arabic, 8 French, 10 Korean, 9 Spanish, 9 Chinese, and 1 examinee was Japanese, Nepali, and Greek, respectively. Efforts were made to ensure that approximately half of the Indian examinees speak a major Northern Indian language (Hindi) and the other half speak a major Southern Indian language (Telugu and Tamil) in each scoring session.

Examinees' responses had been scored by TOEFL iBT raters, and the distributions of scores were roughly consistent across tasks and scoring sessions. For each of the three tasks within a scoring session, there were about 30% to 40% responses each at score levels 2 and 3, and 10% to 20% responses each at score levels 1 and 4. Actual frequency distributions of scores varied slightly, due to efforts to achieve balanced L1 representation across scoring sessions (see Appendices A and B).

Training and Scoring Procedures

A special downloadable computer scoring program designed for the study was used to replicate the ETS Online Scoring Network (OSN) used in operational scoring. All task materials (stimulus materials and prompts, benchmark responses, and task support materials) and responses were easily accessible to raters through this program.

Below is a brief description of the regular and special training and scoring procedures employed in this study. The training and scoring procedures for operational scoring are also reviewed here for comparison purposes.

Training and scoring procedures for operational scoring. In a standard TOEFL iBT Speaking scoring session, raters train and calibrate before scoring each task type. Before raters begin to score each speaking task type, they must first review benchmark responses along with rationales for the assigned scores and then take and pass a short calibration test. The benchmark and calibration responses are generic training materials; they are based on the same *task type* to be scored, but of different *task content*. Benchmark and calibration responses consist of a representative sample of examinees from various first-language backgrounds at the four proficiency levels represented in the scoring rubrics. After raters have reviewed all of the training responses for a task type, and passed calibration, they may then go on to score that task. In a standard scoring session, raters receive guidance from a scoring leader, former raters who have demonstrated consistent scoring accuracy as well as knowledge of and adherence to scoring policies and procedures. Each scoring leader is responsible for 8 to 10 raters, monitoring their scoring performance and mentoring by telephone or e-mail.

Regular training and scoring procedures for this study. In Scoring Session 1, raters followed standard training procedures, first scoring Task 2. Raters reviewed the generic benchmarks for Task 2. The researchers, acting as scoring leaders, provided guidance, commenting on salient features of the responses and the relationship to the scoring guide descriptors. Calibration responses were played while raters scored the responses independently. Scores were checked, and where there was variation within the group, rationales for pre-assigned scores were discussed. Raters then independently scored Task 2 responses for approximately 2.5 hours. Benchmark responses and topic support materials were available for review throughout the scoring session. After scoring for Task 2 was completed, identical processes were followed for training and scoring Tasks 4 and 5 consecutively.

Special training and scoring procedures for this study. During Scoring Session 2, Group 1 continued to follow regular training procedures for the second round of scoring. Group 2 received special training in addition to access to the regular training materials.

Scoring began with a new set of 100 responses for Task 2. While Group 1 reviewed the same generic mixed-L1 benchmark and calibration responses used in Scoring Session 1, Group 2 reviewed a special set of benchmark responses consisting of Indian examinees only. Salient features of the benchmark responses and rationales for the scores were discussed by the group. In addition to the special benchmark samples, the generic benchmark responses used in the regular training in Scoring Session 1 were also reviewed by the raters on their own. Group 2 then scored a special calibration set consisting of only Indian speaker responses. Following Task 2 training, each group scored responses for approximately 2.5 hours. Groups then trained for and scored Tasks 4 and 5 consecutively with Group 1, the control group, following standard training practices and Group 2, the experimental group, following the special training procedures.

Rater Feedback Surveys

Two rater feedback surveys were designed, one completed by all raters at the end of Scoring Session 1 and the second completed by raters in the special training group after scoring Session 2. The first survey included questions about the adequacy of the online training tutorial and certification test for preparing the raters for scoring. It also asked about the challenges they experienced in scoring Indian examinees and their confidence in scoring them. In addition, information about their exposure to non-Indian accents and difficulty in scoring non-Indian examinees was solicited. The second survey focused on the effectiveness of the special training and the overall confidence of the raters in scoring Indian examinees after going through the special training.

Data Analyses

The agreements between raters were estimated using percentages of perfect, adjacent and nonadjacent agreements, quadratically weighted kappa, and Pearson correlations. These different agreement indices provide complementary information about the level of rater agreement observed. While exact, adjacent, and nonadjacent agreements between raters are routinely reported in the score reliability studies involving performance assessments, they are susceptible to chance agreement between raters. Kappa (Cohen, 1960) is a commonly used coefficient of agreement and is considered more robust than simple percent agreement, since it takes into account the agreement occurring by chance. Unlike kappa, which treats all disagreements equally, weighted kappa (Cohen, 1968) weights different types of disagreements differentially, giving more weight to disagreements of greater gravity. The weights assigned to different types of disagreement reflect the costs associated with them. In this study, the TOEFL iBT Speaking responses were scored in the range of 1-4. In operational scoring, although perfect agreements are desired, adjacent agreements between raters are acceptable (e.g., 2 vs. 3) and are not adjudicated. However, discrepancies of 2 or more points are considered serious disagreements and adjudicated. Accordingly, discrepancies of 2 or more points should be heavily penalized in computing the coefficient of agreement. Therefore, quadratic weights are used in this study (e.g., 0 assigned to perfect agreement, 1 to adjacent agreement, 4 to discrepancy of 2 points, and 9 to discrepancies of 3 points).

Correlation is a metric that looks at how consistently two raters rank order examinees, not taking account of consistent differences in the raters' mean scores. It forms the basis for computing inter-rater reliability that is commonly reported in many rater reliability studies.

In addition, the reliability of scores was estimated using G theory. Mean ratings across raters on relevant questions in the rater surveys were computed.

Results

Question 1: How Did the Raters From India Perform in Scoring the TOEFL iBT Speaking Section With Training Similar to That Received by Operational ETS Raters?

The scores assigned by raters from India and ETS raters were compared. Table 3 presents the average agreements between scores assigned by the raters from India and by operational ETS raters. The agreements were calculated in a few different ways: proportion of exact agreement, proportion of exact plus adjacent agreement, quadratically weighted kappa, and correlation. The agreements at both the item level and the aggregated score level (summed across three items) are shown. The exact agreement rate, exact and adjacent agreement rate, correlation, and kappa were averages across all estimates between each Indian rater's scores and the ETS official scores. The distributions of the agreements (average, maximum, minimum, and standard deviation) are provided in Appendix C. Overall, there was a small amount of variation in the agreements with the ETS rater scores across the 26 raters from India.

It has to be noted that in the operational scoring of TOEFL iBT Speaking, a proportion of the responses are double-scored to monitor rater performance. Cases where the two scores differ by 2 points or more are adjudicated by a scoring leader. Thus ETS raters' scores were either ETS Rater 1 scores or adjudicated scores (a few out of 600 cases).

It is meaningful to compare the agreements between the scores of raters from India and ETS raters and those between two ETS raters. Because only a small portion of the scoring samples used in this study were double-scored by ETS raters, the rater agreements were computed on the double-scored responses for the entire test administration. Table 4 contains the

Table 3

Agreements Between Indian Rater Scores and ETS Rater Scores in Scoring Session 1 With Regular Training

	Exact agreement [a]	Exact & adjacent agreement	Weighted kappa	Correlation
All responses (*N* = 300, *M* = 2.52, *SD* = 0.92)	58.4%	97.9%	.74	.75
Indian responses (*N* = 150, *M* = 2.62, *SD* = 0.88)	56.2%	97.7%	.69	.71
Non-Indian responses (*N* = 150, *M* = 2.42, *SD* = 0.96)	60.6%	98.1%	.77	.78
Item 2	59.5%	98.0%	.74	.76
Item 4	56.6%	97.4%	.74	.75
Item 5	59.1%	98.2%	.74	.76
All responses (scores summed across three tasks)	N/A	N/A	.87	.88
Indian responses (scores summed across three tasks)	N/A	N/A	.82	.86
Non-Indian responses (scores summed across three tasks)	N/A	N/A	.90	.92

[a] The exact agreement rate, exact and adjacent agreement rate, kappa, and correlation were averaged across all estimates between each Indian rater's scores and the ETS official scores.

agreements between raters from India and ETS raters and between ETS raters. Comparing the kappa values and correlation estimates in Table 4, it may appear that the agreement between the scores of the raters from India and ETS Rater 1 scores was higher than that between two ETS raters. However, this was due to differences in the score distributions of the scoring sample and the entire test administration. In particular, the scores were more evenly distributed across the four score levels and were more varied in the scoring sample used for Scoring Session 1. Using the algorithm described in Chapter 9 of Haberman (1979), we adjusted the correlations and kappas between ETS raters' scores with the marginal distributions of the scoring sample. The results show that the correlations and kappas would be largely similar to those observed between raters from India and ETS Rater 1 scores, if adjusted with marginal distributions similar to those of the scoring sample. There were two noticeable differences: on Task 4, the ETS rating pairs showed better agreement; on Task 2, the raters from India agreed more with the ETS Rater 1.

As mentioned earlier, those raters who passed the certification test with the lenient criteria and those who passed with the standard criteria participated in the study. Table 5 shows that the scoring performance of the near-pass group was comparable to that of the pass group, as indicated by the agreements with the official scores at the task score level and the aggregated score level.

Table 4

Agreements Between Raters From India and ETS Raters and Between ETS Raters on the Double-Scored Responses for the Entire Test Administration

	Indian Rater – ETS Rater 1		ETS Rater 1 – ETS Rater 2	
	Kappa	Correlation	Kappa	Correlation
All responses (*N* = 963, *M* = 2.53, *SD* = 0.76)	.74	.75	.63 **(.74)**	.63(**.74**)
Indian responses (*N* = 275, *M* = 2.91, *SD* = 0.69)	.69	.71	.53**(.68)**	.53**(.68)**
Non-Indian responses (*N* = 688, *M* = 2.38, *SD* = 0.74)	.77	.78	.62**(.76)**	.62**(.76)**
Item 2 (*N* = 325)	.74	.76	.57(**.71)**	.57**(.71)**
Item 4 (*N* = 314)	.74	.75	.69**(.78)**	.69**(.78)**
Item 5 (*N* = 324)	.74	.76	.62 **(.73)**	.62 **(.73)**

Note. The numbers in boldface are the adjusted correlation and kappa estimates given the marginal totals of the Indian scoring sample used in Scoring Session 1.

In summary, regarding Question 1, we found that the agreement between Indian raters' scores and ETS Rater 1 scores was as high as that between ETS raters. In addition, the near-pass raters showed similar performance as the pass raters after the regular training.

Table 5

Agreements Between Raters From India Scores and ETS Official Scores in Scoring Session 1 With Regular Training—Pass vs. Near-Pass Groups

	Pass group				Near-pass group			
	Exact	Exact & adjacent	Weighted kappa	Corr	Exact	Exact & adjacent	Weighted kappa	Corr
Scoring Session 1 (single task)	58.6%	97.7%	.74	.75	58.2%	98.1%	.74	.75
Scoring Session 1 (three tasks)	N/A	N/A	.87	.88	N/A	N/A	.87	.88
Scoring Session 2 (single task)	58.7%	98.0%	.74	.75	59.4%	97.4%	.74	.75
Scoring Session 2 (three tasks)	N/A	N/A	.87	.88	N/A	N/A	.86	.88

Question 2: To What Extent Did the Special Training Impact the Quality of Scores Assigned by Raters From India?

We addressed this question in a few different ways. First, we compared the agreements between the scores from the raters in India and ETS raters on Indian examinees for the regular training and special training groups. Then we compared the reliability of the scores assigned by the raters in these two groups. We also broke down the analyses by pass and near-pass groups to examine whether the special training gave the less prepared raters a special boost.

Did the special training improve the agreements between raters from India scores and ETS scores on Indian examinees? No noticeable differences were found in the agreements with the ETS raters' scores at the task level across the regular training and special training groups (Table 6). However, when the scores were averaged across the three tasks, the special training group had on average a slightly higher agreement rate with the ETS scores on the Indian examinees in Scoring Session 2 than the regular training group, as indicated by the kappa estimates (.86 and .83, respectively).

Did the special training improve the agreements between near-pass raters and ETS raters on Indian examinees? Again, when the scores were summed across three tasks, the near-pass group who were trained using the special procedures gained an advantage, as both the kappas and correlations with the ETS raters' scores increased after the training (kappa: .79 vs. .84; correlation: .84 vs. .87; see Table 7). By contrast, the near-pass group trained in the regular way kept the same level of agreement with the operational raters across the two scoring sessions (kappa: .85 vs. .84; correlation: .87 vs. .87). However, we also realize that the near-pass raters in the special training group had lower agreement with ETS scores to begin with in Scoring Session 1, so there may have been more room for improvement for this group.

Table 6

Agreement With the Official Scores on Indian Examinees—Regular vs. Special Training Groups[3]

	Regular training group				Special training group			
	Exact agreement	Exact & adjacent agreement	Kappa	Corr	Exact agreement	Exact & adjacent agreement	Kappa	Corr
Scoring Session 1 (single task)	56.5%	97.7%	.69	.71	55.8%	97.6%	.69	.71
Scoring Session 2 (single task)	57.4%	97.8%	.68	.70	57.6%	98.1%	.69	.71
Scoring Session 1 (three tasks)	N/A	N/A	.82	.86	N/A	N/A	.82	.86
Scoring Session 2 (three tasks)	N/A	N/A	.83	.87	N/A	N/A	.86	.88

Table 7

Agreement Between Raters From India and ETS Raters on Indian Examinees—Pass vs. Near-Pass Raters

	Regular training group				Special training group			
	Pass		Near-pass		Pass		Near-pass	
	Kappa	*r*	Kappa	*r*	Kappa	*r*	Kappa	*r*
Scoring Session 1 (single task)	.67	.70	.71	.72	.71	.73	.66	.69
Scoring Session 2 (single task)	.67	.69	.70	.72	.72	.73	.67	.69
Scoring Session 1 (three tasks)	.80	.85	.85	.87	.84	.87	.79	.84
Scoring Session 2 (three tasks)	.83	.87	.84	.87	.87	.89	.84	.87

Did the special training help the raters score Indian examinees more consistently? To answer this question, we conducted four fully crossed G studies (person by rater by task) that corresponded to the four cells in Table 8. The phi coefficients, which are reliability estimates for absolute decisions (i.e., reliability estimates that concern the precision of the scores students receive rather than the rank ordering of them), are reported. Specifically, the phi coefficients for single scores combined across six tasks are reported in Table 8. Only three of the six tasks in a TOEFL iBT Speaking section were scored in this study due to time constraints. The variance components estimated based on the three tasks (G study) were used to project the reliability of the scores averaged across six tasks (Decision, or D study).

Table 8

Phi Coefficients for Single Scores of Indian Examinees Combined Across Six Tasks for Regular vs. Special Training Groups[4]

	Phi coefficient for single score, six tasks	
	Regular training group	Special training group
Scoring Session 1	.84	.85
Scoring Session 2	.84	.90

Before the special training was carried out, the scores assigned by the two rater groups on the Indian examinees were equally reliable, having almost the same phi coefficient for single scores combined across six tasks (.84 and .85). This is to say, if we randomly selected a rater to score each response from the regular training group or the special training group, the reliability of the total test scores (scores averaged across six tasks) would be .84 and .85, respectively. However, the scores assigned by the special training group to the Indian examinees had a much higher phi coefficient (.90) during Scoring Session 2 than the regular training group (.84).

A close examination of the variance components associated with different sources of variation pinpointed the areas that the special training impacted. Table 9 compares the variance components associated with various effects, that is, sources of variation for the regular training group versus the special training group during Scoring Session 2. As is shown, all the variance components involving raters (*r, pr, rt,* and *prt & random error*) were smaller for the special training group. This suggests that the raters in the special training group were more similar in exercising their leniency or harshness and more consistent in rank-ordering Indian examinees than those in the regular training group.

A comparison of the contributions to score variation attributable to different sources for the special training group under the two training conditions shows that all the variance components associated with raters were reduced in Scoring Session 2 (Table 10).

Table 9

Variance Components for Regular Training Group vs. Special Training Group (Scoring Session 2)

Sources of variation	Regular training group		Special training group	
	Variance component	Percent of total variation	Variance component	Percent of total variation
p	0.540	61.2	0.604	68.0
r	0.032	3.6	0.013	1.5
t	0.000[1]	0.0	0.001	0.1
pr	0.021	2.4	0.011	1.2
pt	0.059	6.6	0.063	7.0
rt	0.016	1.8	0.011	1.3
prt & random error	0.215	24.3	0.186	20.9

Note. Rounded off to three decimal places.

Table 10

Variance Components for the Special Training Group in Scoring Session 1 and Scoring Session 2

Sources of variation	Scoring Session 1		Scoring Session 2	
	Variance component	Percent of total variation	Variance component	Percent of total variation
p	0.494	58.9	0.604	67.9
r	0.018	2.1	0.013	1.5
t	0.000[1]	0.0	0.001	0.1
pr	0.020	2.4	0.011	1.2
pt	0.082	9.8	0.063	7.1
rt	0.020	2.4	0.011	1.2
prt & random error	0.204	24.3	0.186	20.9

Note. Rounded off to three decimal places.

Did the special training improve the reliability of the scores of Indian examinees assigned by the near-pass raters? The G study analyses reveal that the near-pass raters under the special training were able to rate more consistently, as the phi coefficient increased considerably (from .79 to .89) (see Table 11). In contrast, the reliability of the scores for the near-pass raters who went through regular training during both scoring sessions showed almost no change across the two sessions (.86 vs. .85). Looking at the variance components associated with near-pass raters (Table 12), the person-by-rater interaction reduced considerably after the special training, and the rater main effect and the rater-by-task interaction decreased slightly as well. These results point to the conclusion that the near-pass raters assign much more similar scores when scoring Indian examinees when they receive special training than when they do not.

Table 11

Phi Coefficients for Single Scores Combined Across Six Tasks on Indian Responses—Pass vs. Near-pass Groups

	Phi coefficients for single score & six tasks			
	Regular training group		Special training group	
	Pass group	Near-pass group	Pass group	Near-pass group
Scoring Session 1	.83	.86	.87	.79
Scoring Session 2	.83	.85	.91	.89

Table 12

Variance Components for the Near-pass Raters in the Special Training Group in Scoring Session 1 and Scoring Session 2

Sources of variation	Scoring Session 1		Scoring Session 2	
	Variance component	Percent of total variation	Variance component	Percent of total variation
P	0.441	54.2%	0.590	64.3%
R	0.020	2.5%	0.016	1.7%
T	0.000[1]	0.0%	0.012	1.3%
Pr	0.043	5.3%	0.008	0.9%
Pt	0.102	12.5%	0.067	7.3%
Rt	0.020	2.5%	0.010	1.1%
prt & random error	0.187	23.0%	0.214	23.3%

Note. Rounded off to three decimal places

Question 3: How Did the Raters From India Perceive Their Scoring Experiences at the End of Scoring Session 1?

Table 13 provides the averaged ratings of the raters to a few selected questions in the survey administered at the end of Scoring Session 1. In general, these raters felt the online training tutorial and the certification test prepared them well for scoring TOEFL iBT Speaking. They did not think that their familiarity with Indian accents created much difficulty for them to evaluate Indian examinees fairly and accurately. They felt equally confident in scoring Indian and non-Indian examinees.

Table 13

Average Ratings on Selected Questions in the Survey Given at the End of Scoring Session 1

Survey question	Average rating (*n* = 26)
Would completing the online training tutorial and passing the certification test make you feel ready to score TOEFL iBT Speaking responses?	3.6 (not ready–sufficiently ready) (Min = 2, Max = 4, SD = 0.6)
Did your familiarity with the accent make it difficult to provide a fair and accurate evaluation of Indian speakers' speaking proficiency?	1.6 (not difficult–very difficult) (Min = 1, Max = 3, SD = 0.9)
How would you rate your overall confidence level in scoring Indian speakers?	3.6 (not at all confident–very confident) (Min = 3, Max = 4, SD = 0.5)
How would you rate your overall confidence level in scoring non-Indian speakers?	3.6 (not at all confident–very confident) (Min = 3, Max = 4, SD = 0.5)

Although the raters reported being highly confident in scoring Indian examinees in this context, when asked about some specific scoring challenges they had, some felt uncertain about the impact of their greater exposure to Indian accents on their scoring. They commented that because they were more used to the accents, they were not sure if they were too lenient or too harsh while trying to correct for their familiarity. They also noted that the rapid-fire delivery characteristic of many Indian examinees created difficulty in scoring because it gave an illusion of fluency; so in scoring they needed to listen attentively to content. The nonflat profiles of some Indian examinees (i.e., showing good content and vocabulary but falling short on delivery) also seemed to have made it harder for them to make holistic judgments of proficiency.

When asked to suggest rater training and support materials that would help them score Indian examinees, the raters expressed the need for more detailed guidelines for evaluating the pronunciation and intonation patterns of Indian examinees. They also felt that it would be useful to demonstrate the levels of Indian accents that are comprehensible to native speakers of English. In addition, they recommended providing responses of Indian examinees as benchmarks, which we did for the special training group in Scoring Session 2.

Regarding specific challenges they had in scoring non-Indian examinees, they indicated that they had some difficulty understanding certain accents and suggested that more practice materials that could familiarize them with different varieties of accents would be useful.

Question 4: How Did the Special Training Group Perceive the Effectiveness of the Special Training?

At the end of Scoring Session 2, the raters who participated in the special training were given a survey on the effectiveness of the training. Their survey responses generally show that they perceived the special training as useful and that it boosted their confidence in scoring Indian examinees. When asked whether reviewing the exemplary responses from Indian examinees was useful in helping them score Indian examinees, they unanimously gave a rating of 4 (very useful) on a scale of 1-4. When asked whether they felt more confident scoring Indian examinees after reviewing exemplary responses from Indian examinees, their average rating was 3.5 out of 4 (Min = 1, Max = 4, SD = 0.9), with 4 indicating much more confident.

Discussion and Conclusion

This study investigated the impact of the familiarity of raters from India with Indian accents on their scoring of Indian examinees’ responses to the TOEFL iBT Speaking section and the effectiveness of a special training package in helping them score Indian examinees. In addition, this study examined these raters’ scoring performance on non-Indian examinees.

Under the training and scoring conditions in this study (i.e., on-site training was conducted and all raters worked at a central location and were paced), the raters from India performed as well as ETS operational raters, if not better. The near-pass raters also emulated the performance of the pass raters, having comparable agreements with the operational raters.

The special training did not give the intervention group an advantage when their scores were compared with the ETS scores at the task level; however, when the scores were summed across the three tasks, the special training group had slightly better agreements with the ETS raters. Furthermore, the special training seems to have made the raters in the special training group more consistent in scoring Indian examinees than those in the regular training group. In other words, their scores on Indian examinees were more alike than those assigned by raters in the regular training group. The special training seemed to have given the near-pass raters a special boost, leading to greater improvement in their scoring consistency than the pass raters who received the special training. The raters in the regular training group also assigned fairly similar scores to Indian examinees, although their scoring consistency was not on a par with that of the special training group. The reasonably good scoring consistency of the regular training group may be attributable to their ESL/EFL training background, which may have helped them develop an ear for discerning different degrees of Indian accents. Their linguistic background in Indian languages may have facilitated their understanding of accented speech of Indian examinees but may not necessarily have led to higher scores than were warranted. Nonetheless, the special training does seem to have enhanced raters' ability to score Indian examinees in a consistent way, because their scores on Indian examinees were more alike after the special training.

Overall, this study demonstrates that with rigorous training and certification procedures, the potential bias associated with a rater's greater familiarity with a particular accent may be minimized. This potential bias could be positive (i.e., assigning higher scores than actually deserved) or negative. Positive bias may result from raters' greater facility in understanding a particular accent. Negative bias may be associated with raters' attempts to overcorrect their tendency to give higher scores or to unfairly penalize examinees who share the same L1 with them because of the long and strenuous process of learning the target language they themselves have gone through. Although both types of scoring tendencies have been reported in studies that employ naive, untrained raters, they can be minimized by adequate rater training practices.

In holistic scoring, raters typically engage in impressionistic judgment of overall speech quality. However, in a situation where raters have to judge the speaking proficiency of examinees who share the same L1 with them, they may need to adjust their typical scoring behavior. In the case of the present study, if raters from India had relied on their first impressionistic judgments of the comprehensibility of Indian examinees, bias could have occurred, since they tend to understand Indian examinees better than operational raters. To provide accurate and consistent evaluations of Indian examinees, they may have resorted to more analytic evaluations. The raters trained with the special training package may have used the

Indian benchmarks and calibration samples as their guide to determine how similar a response was to exemplary responses at a particular score level. The exemplary samples may have helped them reinforce their impressions of Indian examinees at different score levels irrespective of how much they could understand a speaker and how much listener effort is involved. Although the raters trained using the regular procedures were not able to benefit from the Indian benchmarks and calibration samples, their years of experience teaching English to Indian examinees may have helped them develop a certain degree of competence in recognizing different degrees of accents. Our hypothesis was that the raters from India may have engaged in more analytic evaluations in scoring Indian examinees, while relying on more impressionistic, intuitive judgments for examinees whose L1s were not familiar to them. However, this speculation needs to be corroborated by direct empirical evidence, such as rater verbal protocol data.

Regarding the perceived effectiveness of the special training, the raters in the special training group perceived the training as very helpful and felt more confident in scoring Indian examinees after the training. This demonstrates that the special training had positive psychological effects on the raters, which paralleled improvements in the agreement between their scores and the ETS scores and in their scoring consistency.

This study did not involve rater verbal protocols, although rater protocol analyses, along with the survey data we collected, would shed more light on the specific challenges the raters from India experienced in scoring, differences in their scoring behaviors associated with Indian and non-Indian examinees, and how exactly the special training package may have helped them evaluate Indian examinees.

This study did not explore how raters’ familiarity with examinees’ L1 impacts their evaluations of specific aspects of speech, such as pronunciation, intonation, or stress patterns. While this study showed no impact on raters’ holistic judgments, there may well be differences in how they evaluate some analytic components of speech.

One point worth mentioning is that for the generalizability analyses performed in this study, the universe of raters (the pool of raters to which the results can generalize) were those who have similar qualifications and exhibit similar scoring competence as the raters from India used in this study. Because our purpose was to examine the possibility of including raters from India in the rater pool for operational scoring, strictly speaking, generalizability studies that include both a sample of operational raters and raters from India (the combined group would be the universe of raters) would provide more conclusive evidence about the impact of including raters from India on the score reliability. We were not able to include a sample of operational raters due to practical constraints. Nevertheless, the results of this study warrant the following conclusions: First, the raters from India in this study were comparable to the operational TOEFL iBT raters, as evidenced in the high agreements between their scores and the operational scores. This suggests that these raters from India can be considered interchangeable with the operational raters. Second, although we did not model both groups in a single G theory analysis, given the comparability between the raters from India and operational raters, the generalizability analyses based on the scores assigned by the raters from India show, albeit in an indirect way, that their scores would have a high level of reliability.

References

Barnwell, D. (1989). “Naïve” native speakers and judgments of oral proficiency in Spanish. *Language Testing, 6*(2), 152-63.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.

Brodkey, D. (1972). Dictation as a measure of mutual intelligibility. *Language Learning, 22*(2), 203-220.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1), 1-15.

Caban, H. L. (2003). Rater bias in the speaking assessment of four L1 Japanese ESL. *Second Language Studies, 21*(2), 1-44.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16-33.

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes, 24*(3), 383-391.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Cohen J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70,* 213-20.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scales and profiles*. New York: Wiley.

Davies, A., Hamp-Lyons, L., & Kemp, C. (2003) Whose norms? International proficiency tests in English. *World Englishes, 22*(4), 571-584.

Eckes, T. (2005). Examining rater effects in TestDaF Writing and Speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*(3), 197-221.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37-64.

Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work*? Language Assessment Quarterly, 2*(3), 175-196.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning, 37*(3), 313-26.

Galloway, V. (1980). Perceptions of the communication efforts of American students of Spanish. *Modern Language Journal, 64*(4), 428-33.

Gass, S. M., & Varonis, E. (1984). The effect of familiarity on the comprehension of non-native speakers. *Studies in Second Language Acquisition, 7*(1), 37-57.

Gorosch, M. (1973). Assessment intervariability in testing oral performance of adult students. In J. Svartvik (Ed.), *Errata: Papers in error analysis* (pp. 145-153). Gleerup Publishers: Lund, Sweden.

Haberman, S. J. (1979). *Analysis of qualitative data: Vol 2, new developments*. New York: Academic Press.

Hadden, B. L. (1991). Teacher and nonteacher perceptions of second language communication. *Language Learning, 41*(1), 1-24.

Indopedia. (2004). *Indian English*. Retrieved June 10, 2009, from http://www.indopedia.org/Indian_English.html

Kachru, B. B. (1986). *The alchemy of English: The spread, functions and models of non-native Englishes*. Oxford, England: Pergamon.

Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187-217.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26-43.

Major, R., Fitzmaurice, S. F., Bunta, F., & Balasubramanian, C. (2002). The effects of nonnative accents on listening comprehension: Implications for ESL assessment. *TESOL Quarterly, 36*(2), 173-190.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of L2 speech. *Studies in Second Language Acquisition, 28*(1), 111-131.

Quirk, R. (1985). The English language in a global context. In R. Quirk & H. G. Widdowson (Eds.), *English in the world: Teaching and learning the language and literatures* (pp. 1-6). Cambridge, England: Cambridge University Press.

Quirk, R. (1990). Language varieties and standard language. *English Today, 21*, 3-10.

Sheorey, R. (1985). *Goof gravity in ESL: Native vs. nonnative perceptions.* Paper presented at the 19th annual convention of Teaching English to Speakers of Other Languages (TESOL), New York.

Shohamy, E., Gordon, C.M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal, 76*(1), 27-33.

Smith, L. E., & Bisazza, J. A. (1982). The comprehensibility of three varieties of English for college students in seven countries. *Language Learning, 13*(2), 259-269.

Smith, L. E., & Rafiqzad, K. (1979). English for cross-cultural communication: The question of intelligibility. *TESOL Quarterly, 13*(3), 371-380.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.

Xi, X., & Mollaun, P. (2006). *Investigating the utility of analytic scoring for the TOEFL Academic Speaking Test (TAST)* (TOEFL iBT Research Rep. No. TOEFLiBT-01). Princeton, NJ: ETS.

Notes

[1] In MFRM, rater bias refers to systematic patterns of leniency or harshness associated with particular examinee groups, task types, or scoring criteria.

[2] The certification criterion was determined based on a pilot study using potential raters prior to the launch of the TOEFL iBT test. The agreement rate (70% exact agreement, no discrepant scores) is also consistent with what is typically observed on monitor responses (i.e., responses with pre-assigned scores that are used to monitor raters' performance) in operational scoring.

[3] Similar agreement was found on non-Indian examinees across Scoring Session 1 and Scoring Session 2 for the regular training group and the special training group.

[4] Similar phi coefficients were found for single scores of non-Indian examinees combined across six tasks for regular training vs. special training groups across Scoring Session 1 and Scoring Session 2.

Appendix A

Distribution Statistics of the Scoring Samples in Scoring Session 1

Table A1

Percentage of Responses at Each Score Level by Item (All Native Languages)

	Score			
Item	1	2	3	4
2	13%	35%	36%	16%
4	17%	35%	32%	16%
5	13%	35%	36%	16%

Table A2

Percentage of Responses at Each Score Level by Item (Indian Examinees Only)

	Score			
Item	1	2	3	4
2	6%	38%	38%	18%
4	14%	36%	32%	18%
5	8%	36%	40%	16%

Table A3

Percentage of Responses at Each Score Level by Item (Non-Indian Examinees Only)

	Score			
Item	1	2	3	4
2	20%	32%	34%	14%
4	20%	34%	34%	12%
5	18%	34%	32%	16%

Appendix B

Distribution Statistics of the Scoring Samples in Scoring Session 2

Table B1

Percentage of Responses at Each Score Level by Item (All Native Languages)

	Score			
Item	1	2	3	4
2	14%	42%	29%	15%
4	12%	32%	41%	15%
5	12%	36%	37%	15%

Table B2

Percentage of Responses at Each Score Level by Item (Indian Examinees Only)

	Score			
Item	1	2	3	4
2	8%	44%	33%	15%
4	2%	33%	50%	15%
5	8%	38%	35%	19%

Table B3

Percentage of Responses at Each Score Level by Item (Non-Indian Examinees Only)

	Score			
Item	1	2	3	4
2	21%	40%	25%	15%
4	23%	31%	31%	15%
5	17%	33%	40%	10%

Appendix C

Distributions of the Agreements Between Scores of ETS Raters and Raters From India in Scoring Session 1 With Regular Training

		Exact	Exact + adjacent	Weighted kappa	Pearson
All responses	Avg	0.58	0.98	0.74	0.75
	Max	0.65	0.99	0.78	0.79
	Min	0.51	0.96	0.70	0.72
	SD	0.03	0.01	0.02	0.02
Indian responses	Avg	0.56	0.98	0.69	0.71
	Max	0.66	0.99	0.75	0.76
	Min	0.46	0.94	0.62	0.66
	SD	0.05	0.02	0.04	0.03
Non-Indian responses	Avg	0.61	0.98	0.77	0.78
	Max	0.66	1.00	0.82	0.83
	Min	0.51	0.95	0.72	0.74
	SD	0.04	0.01	0.03	0.03

Test of English as a Foreign Language
PO Box 6155
Princeton, NJ 08541-6155
USA

To obtain more information about TOEFL
programs and services, use one of the following:

Phone: 1-877-863-3546
(US, US Territories*, and Canada)

1-609-771-7100
(all other locations)

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

*America Samoa, Guam, Puerto Rico, and US Virgin Islands