ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 5, No. 2, pp. 313-322, March 2014
© 2014 ACADEMY PUBLISHER Manufactured in Finland.
doi:10.4304/jltr.5.2.313-322

College EFL Teachers' Perspectives on Listening Assessment and Summarization for a Specific Task

Brent G. Walters
Chung Yuan Christian University, Taiwan

Ching-ning Chien (Corresponding Author)
Chung Yuan Christian University, Taiwan

***Abstract*—Listening assessment has been a neglected area of research and teaching in the past due to difficulties such as separating input comprehension from output ability, rater variability, and text ambiguity. The goal of this study was to develop a norm-referenced and holistic model summary by surveying Taiwanese English listening and speaking teachers and then comparing that model summary to listening summaries created by non-expert college age native speakers. Eleven Taiwanese college English listening and speaking teachers were surveyed to determine i) their general listening assessment preferences and ii) how they would score a specific listening task, including selecting the key main ideas, key vocabulary, and drafting a model summary. For comparison purposes, the listening task was administered to 10 college-age native speakers of English and they were asked to orally construct a summary of the listening task. The results showed that the teachers were consistent in their listening assessment preferences and also in their preferred assessment choices for a specific listening task. The native-speaker produced summaries did not converge on a single model for the same listening task.**

***Index Terms*—listening assessment, listening comprehension, teacher assessment practices, summarization, native speakers**

I. Introduction

The research into listening assessment practices of language teachers, not just Taiwanese language teachers, is scarce (Alderson & Banerjee, 2002; Chen & Tsai, 2012). Even more scarce are studies investigating the construction of L2 listening summaries, as research into L2 reading summaries is only beginning to be developed (Frost, Elder, & Wigglesworth, 2012; Yu, 2013). As much of the research and practice surrounding L2 listening is derived from L2 reading, many of the most commonly adopted ways to score the listeners' understanding are modelled on assessment forms refined in the reading area such as multiple choice tests, cloze tests, completion (or dictation), and short answer questions. However, the ability to process input by constructing mental (or even written) summaries is critical for successful communication, in school and out, and summary construction has long been seen as "crucial for education" across the curriculum (Seidlhofer, 1995). For listening assessment purposes, summarizing goes beyond mere comprehension because it reflects the listener's identification and reconstruction of the most important elements of the text through his or her analysis and synthesis. In this way, summarization as assessment is a much more robust measure of understanding than multiple choice, cloze, dictation, or even short answer questions. However, the research into summaries has long been complicated by the issue of validation, namely, how to distinguish good summaries from poor summaries (Frost, Elder, & Wigglesworth, 2012). As a result, the use of summaries in listening assessment is not established well enough to practically guide EFL teachers to make use of model listening summaries in their teaching practices. These two areas—Taiwanese listening and speaking teachers' assessment practices and the construction of 'model' listening summaries—are at the focal point of this study.

a. Summaries for assessment

For advanced listening training, summarization is one of the best teaching methods. Summarization is a high-skill exercise, which requires listeners to re-organize the ideas that have been formed while listening (Kirkland & Saunders, 1991). Deciding which information should be included in a summary and how the summary should be organized are normative claims, outside the bounds of the common definition, which must be answered on a case-by-case basis, and are dependent on the context, audience, and purpose of the summary (Seidlhofer, 1995). From this, the two most consistent criticisms against using summaries to assess language ability have been (1) that numerically scoring or differentiating a good summary from a poor summary is too difficult to be practical and (2) that students may not be able to produce output in the L2 commensurate with their ability to comprehend L2 input (Alderson, 2000, pp. 232-233; Frost, Elder, & Wigglesworth, 2012; Kirkland & Saunders, 1991; Yu, 2007). This may explain why summaries, as a form of language assessment, have been so neglected, even though their potential, as a comprehensive and integrative assessment tool, is so high.

Notwithstanding that the research into the use of summaries in L2 pedagogy is underdeveloped, there have been a few notable studies which are important for this research. Rost (1994) compared L2 student summaries, completed after listening to a lecture, to expert native speaker summaries, to identify ways to improve student lecture understanding. Rost, using content and style analysis, found that the lecture was probably too difficult for the L2 learners involved in the study and that they tended to repeat or remodel chunks and phrases from the original lecture without fully grasping the meaning of their new constructions. Keck (2006), comparing L1 and L2 writers' use of paraphrase in the construction of summaries, similarly found that L2 writers use more "near copies" than L1 writers.

Yu (2007) completed a postmodern reading-writing assessment study, comparing the rating scales of native speaking expert scorers with scales constructed by the L2 learners themselves. Although both rating scales were effectively norm-referenced, L2 learners' summaries rated higher under their own "democratically" designed rating systems than they did under the expert native speaker rating systems. Additionally, the expert-designed rating system was not significantly (statistically) better at predicting student achievement on independent measures of student ability such as the First Certification in English "FCE" and TOEFL even though the rating systems only overlapped in their "required key points" by 50%. Furthermore, although the majority of the Chinese L2 learners involved in the study preferred expert-designed rating scales, a substantial number of students thought the democratically derived rating systems were more fair because "students could have their own unique understanding of a text, which 'old' experts might not fully appreciate due to a generation gap" (p. 555).

Inoue (2009) constructed an empirical, discoursed-based rating system by reviewing oral interviews of 12 L2 learners whereby they had to complete numerous tasks including oral summarization. Inoue, in line with previous research, found that the higher proficiency L2 learners paid more attention to details than lower proficiency L2 learners, mimicking the differences found between L2 learners and native speakers when performing summarization tasks. Frost, Elder, and Wigglesworth (2012) had a similar finding in that higher proficiency L2 learners included more details and had a better schematic structure to their summaries than lower proficiency learners.

Finally, Vorobel and Kim (2011) qualitatively analyzed the summaries of a diverse group of seven L2 students studying in the US. They showed the impact that personal factors, such as cultural thinking patterns, motivation, content knowledge, literacy skills, and vocabulary, have on summary construction. Vorobel and Kim advised teachers to be mindful of students' background in order to provide better instruction when dealing with summaries.

b. Difficulties in Rating Summaries

As the previous studies demonstrate, the difficulty in using summaries in language assessment is that the scoring system is not easily validated and reliability can be problematic. Traditionally, summaries were scored holistically or impressionistically, with a teacher selecting, *a priori*, the most important elements required for an "adequate" summary (Frost, Elder, & Wigglesworth, 2012; Seidlhofer, 1995). This presumes that teachers are capable of producing valid and reliable summary rubrics, something which even experts have trouble with (Cohen, 1993, p. 137). Furthermore, the fact that the definition of a "summary" may be disputed, as Cohen (1993) discovered that test takers and raters disagreed over what exactly was a "main idea" and whether or not personal commentary or general knowledge should be included in the summary, is just a reflection that a summary is dependent on components such as main ideas and key vocabulary, each with their own level of ambiguity (Seidlhofer, 1995). This inherent ambiguity is reflected in Yu's work showing a sharp distinction between what students and experts think is a valid summary (2007). Other technical explorations into summary assessment have focused on discourse analysis such as that developed by Seidlhofer (1995), Frost, Elder, and Wigglesworth (2012), and most recently by Yu (2013). Although studies such as these are making worthwhile contributions to the field generally, their elaborate and intensive methodology makes them more suited for diagnostic purposes than regular classroom use.

The difficulty in scoring summaries is compounded by the trend towards incorporating more authentic materials in the language learning classroom: since such materials are designed for native speakers, the context and subtext make clear or definitive comprehension checks, such as through a summary, even more difficult. From a broader perspective, test designers and teachers must recognize that it may not be possible to effectively disassociate our own background knowledge, assumptions, opinions, and biases to credibly claim that we have constructed a "model summary" (Alderson, 2000, p. 150).

c. Research questions

The motivation behind this study was to develop a rating scale for EFL instructors to assess listening comprehension for a specific listening task. From this, three research questions were derived for this study. First, currently which assessment practices are most frequently used by Taiwanese English listening and speaking teachers? In Taiwan, there has been no major work published surveying the practices of this particular population so any insight would be useful. Second, what degree of listening assessment conformity exists between Taiwanese listening and speaking teachers with respect to a specific listening task? This research question was designed to shed light on the degree of similarity of model summaries produced by practicing teachers.

The third research question was, if Taiwanese listening and speaking teachers do show a high degree of listening assessment conformity for a specific task, how closely do their assessment expectations overlap with native speaker listening performance for the same task? Based on extensive research into English teachers for whom English is an L2, we expect a high degree of similarity between what teachers expect L2 learners to understand and what L1 learners understand for a short and limited listening task (Moussu & Llurda, 2008). Many studies have compared native speaking teachers with non-native speaking teachers with respect to speaking assessment (eg., Kim, 2009; Zhang & Elder, 2011), but listening assessment, specifically compared to native speaker performance, appears to be unaddressed in the literature. Using native speakers as a comparison for the teachers' model summaries has two advantages. First, using native speakers in this study greatly reduces the possibility that the participants' output ability is not commensurate with his or her listening ability, as is the case with many L2 study participants (Frost, Elder, & Wigglesworth, 2012; Rost, 1994). Second, using native speaking participants can offer some insight into the upper range of performance that can be expected from L2 learners. In this way, we can get a better picture of how well the teacher-constructed model summaries correlate with higher proficiency performances (Yu, 2007). Ultimately, we hoped that the results would converge to produce a robust ideal model of comprehension for our specific listening task.

II. Methodology

a. Participants

The participants consisted of 11 college English listening and speaking teachers (the "Teachers") and 10 American native-English speaking college students (the "NS"). The Teachers were all female native Chinese speakers. To protect anonymity, no other demographic information about the Teachers was collected as part of this study.

The NS ranged in age from 18 to 21 years old and five were female. All were enrolled in college at the time, although not at the same institution: five were from a single community college, four were from two comparable state universities, and one was enrolled at a military academy. Six of the NS were majoring in psychology, and the remaining four NS were majoring in criminal justice, fashion design, engineering, and journalism. All had normal hearing ability, none had any proficiency in a second language, and none had ever traveled outside the United States.

b. Instruments

The Teachers were given an online survey consisting of two parts: the first part asked for their listening assessment preferences and the second part asked them to identify how they would assess a specific news text. The survey's assessment preferences section consisted of two questions: the first asked them to estimate how often they used particular forms of assessment. The choices offered in the survey were: i) checking for gist; ii) checking for main ideas; iii) checking for details; iv) checking for summarization; v) checking for vocabulary identification and/or comprehension; vi) checking for accuracy (ie., dictation); and vii) genre identification. These choices were made based upon a review and synthesis of Flowerdew and Miller (2005, pp. 184-188), Macaro (2005, pp. 178-179), and Rost's (2002, pp. 172-173) works. The second question asked the Teachers if they used any other listening assessment forms not already listed in the first question.

The second part of the Teachers' survey involved an authentic news broadcast taken from the BBC. The broadcast first aired on April 10, 2010, and was entitled "Poland in mourning for plane crash victims" (BBC, 2010). The broadcast was 1 minute, 23 seconds in length and was structured in traditional news macrostructure (van Dijk, 1988, pp. 15-18) consisting of a lead, macropropositions, details, and conclusion. Given the topical nature and specialized vocabulary with most regular news broadcasts, the short length of this particular news clip would better facilitate student comprehension and subsequent interpretation, and hopefully reduce the difficulty associated with defining summarization (Seidlhofer, 1995). The limited subject nature of the news clip would also help avoid resorting to computational methods or deep semantic analysis to effectually summarize the text. The reporter spoke in accented British English and interviewed Polish speakers who spoke in Polish-accented English. Teachers were given the link to the BBC website which hosted the videos of the news story so that they could hear and see the content of the text. The transcript of the news article, as was also provided to the Teachers in the survey, can be seen in Table 1.

TABLE 1.
TRANSCRIPT OF NEWS ARTICLE USED IN THE TEACHER'S SURVEY

>>Reporter: Poland is a country of churches, but not enough for all those who wish to mourn today.
>>Reporter: An overflow congregation of thousands gathered in Royal Castle Square to hear the Archbishop of Warsaw read the 96 names of those who died. People who in exile, in the Solidarity movement, and now in government, had been at the heart of Poland's modern history.
>>Polish man: This is a tragedy I think for eh, for the country. It's never happened something like this, anywhere in Europe. So uh, it's extremely, uh, sad uh story for us, especially in this week just after the Easter so it's horrible.
>>Polish woman: The most important people in Poland died so I don't know how we'll, the future of our country look like.
>>Reporter: Lech Kaczynski and his identical twin had been child film stars. They both joined the struggle against communism and rose to become President and Prime Minister, leading a strongly nationalistic government, which irritated both Russia and its European Partners.
>>Reporter: Although at times a divisive figure, there has been an outpouring of sorrow.
>>Reporter: Large crowds built up outside the Presidential Palace, laying flowers and lighting candles to mourn this blow to Poland.

* (>>) represents a speaking pause.

Based on the news story in Table 1, the Teachers were asked two multiple choice questions and two open-ended questions regarding how they would assess students. The first multiple choice question asked for Teachers' preferred assessment type from among eight categories: summarization, main idea check, vocabulary check, check for gist, genre identification, reflection/reaction/response, dictation, and other. The second multiple choice question asked Teachers to choose the key main ideas of the news story (Table 2). For both questions, Teachers had the option of selecting multiple responses.

TABLE 2.
KEY MAIN IDEAS OF THE NEWS STORY

• people are mourning
• the most important people in Poland have died
• Poland is full of churches
• people gathered in the square to hear the archbishop
• people gathered outside the presidential palace
• 96 people died
• a tragedy has struck Poland
• the future is uncertain for Poland
• twin brother child film stars became Polish President and Prime Minister
• the Polish President and Prime Minister irritated neighbouring countries

The third question relating to how the Teachers would assess students asked for "the most important vocabulary for understanding the news text". The final survey question asked Teachers to summarize the news story.

c. Native speaker interviews

The 10 native speaking participants were allowed to listen to the news story twice. Participants listened on headphones, after the volume was adjusted for their comfort. During the first listening participants were asked to think aloud concurrently. The second listening was used to recall the participants' thoughts during the first listening event. The second listening involved pausing the recording at natural breaks within the story (ie., every few sentences). After each pause in the playback, participants were asked to report any additional thoughts they may have had during the think aloud or any new information they may be noticing on the second listening. Immediately at the end of the second listening, participants were asked to summarize what they had just heard. All think aloud, recall, and summarization was recorded and then transcribed for analysis.

In earlier pilot studies of this research, it was found that nearly all NS participants were unfamiliar with the subject of the news story or the recent history of Poland and Polish politics. This meant that the NS would be discovering the information in the news story for the first time during their participation in this study. For the data reported here, all but one of the 10 NS study participants reported being completely unfamiliar with the subject of the news story or Poland's recent history. The data from the one participant who reported "vaguely" recalling hearing the news story when it first aired in April 2010 was retained because his memory did not appreciably affect his performance.

d. Data Analysis

After collecting all survey data and transcribing all interviews, text analysis was performed using text analysis software Concordance.

III. RESULTS

The first part of the Teachers' survey concerned their preferred listening assessment practices. The first survey question asked them how often they used specific forms of listening assessment and the results can be seen in Table 3.

TABLE 3.
TEACHERS' PREFERRED LISTENING ASSESSMENT PRACTICES

Frequency	Never	Rarely	Sometimes	Frequently	Exclusively
Checking for gist			27%	64%	9%
Checking for main ideas			9%	82%	9%
Checking for details			27%	55%	18%
Checking for summarization		27%	36%	27%	9%
Checking for vocabulary ID and/or comprehension			18%	64%	18%
Checking for accuracy (ie., dictation)		9%	45%	45%	
Genre identification		27%	45%	27%	

As can be seen from Table 3 above, checking for main ideas, checking for vocabulary identification and comprehension, checking for details, and checking for gist were used "frequently or exclusively" by 91%, 82%, 73%, and 73% of the Teachers respectively. The second question regarding Teachers' preferred listening assessment practices was open-ended, allowing to the Teachers to suggest any other forms of assessment not mentioned above, but no responses were given.

The second part of the Teachers' survey asked them to design listening assessment for a specific news story. Teachers were first asked which assessment type they would most likely use for this specific news text and the results can be seen in Table 4.

TABLE 4.
ASSESSMENT TYPE REPORTED MOST LIKELY TO BE USED

Assessment Type	Response Frequency (%)
Main idea check	100
Check for gist	73
Genre Identification	64
Summarization	45
Reflection / reaction / response	36
Vocabulary check	36
Dictation (semi or full)	9

The second question for the applied portion of the survey asked the Teachers to identify the key main ideas for the news story and the results can be seen in Table 5.

TABLE 5.
SELECTED KEY MAIN IDEAS OF NEWS STORY

Main Idea	Response Frequency (%)
the most important people in Poland have died	91
a tragedy has struck Poland	82
people are mourning	64
the future is uncertain for Poland	36
people gathered in the square to hear the archbishop	27
people gathered outside the presidential palace	27
the Polish President and Prime Minister irritated neighboring countries	27
96 people died	27
twin brother child film stars became Polish President and Prime Minister	9

The third question on the applied portion of the Teachers' survey asked them to identify the key vocabulary in the news story. The Teachers aggregate key word count (including word phrases such as "prime minister") was 61 with 25 unique words. The results of the most common 12 words (2 or more occurrences) are shown in Table 6.

TABLE 6.
REPORTED MOST IMPORTANT VOCABULARY OF THE NEWS STORY

Word or Word Phrase	Occurrences	Frequency (%)
tragedy	10	16
mourn	8	13
died	6	10
Poland	5	8
people	5	8
important	4	7
sorrow	2	3
blow	2	3
country	2	3
prime minister	2	3
most	2	3
president	2	3

The fourth question of the applied portion of the Teachers' survey asked them to construct a model summary for the news story. A word analysis of the summaries found 221 words (including phrases being counted as one word) and 77 unique words. The summaries are reproduced in Table 7 and the key word analysis (for word occurrences of 3 or more) of the model summaries are shown in Table 8.

TABLE 7.
MODEL SUMMARIES FOR THE NEWS STORY

A tragedy struck Poland and people in Poland are mourning over the death of the most important people of their country.
People in Poland are mourning the loss of 96 lives, including the President and Prime Minister, who are the most important people for their supporters. The two brothers joined the struggle against communism and have been leading a strongly nationalistic government.
People get together to mourn for the death of important people in Poland.
A tragedy has happened in Poland. Large crowds gathered to mourn the death of the most important people in Poland.
Tragedy stroke Poland. Some important people died and people in Poland are mourning for the losses.
People in Poland are mourning for a tragic event that happened in their history.
There were some people in who Poland died, and many people were sad.
People in Poland gathered and mourned for the death of 96 important people and their uncertain future.
A tragedy happened in Poland that made the whole country sad, and people gathered to mourn for the death of the most important people in Poland.
96 people, including the President and Prime Minister died and many Polish people gathered outside the church to mourn for them.
The most important people in Poland, including Lech Kaczynkski and his twin brother, died. Many in Poland are mourning for the tragic loss.
A tragedy struck Poland and people in Poland are mourning over the death of the most important people of their country.

TABLE 8.
WORD ANALYSIS OF THE MODEL SUMMARIES FOR THE NEWS STORY

Word or Word phrase	Occurrences	Frequency (%)
people	18	8
Poland, -ish	16	7
death, dead, died	9	4
mourn, -ing	9	4
important	8	4
tragedy, tragic	6	3
gathered	4	2
happened	3	1
including	3	1
loss, losses	3	1

a. Native Speaker Listening Performance

After transcribing all the NS summaries of the news text, the words were analyzed and 639 total words were found, with 210 unique words. The results of the word analysis of the NS summaries are shown in Table 9. Of the 10 NS summaries, only five appeared to closely follow the summary elements contained in the Teachers' model summaries. Due to their length, only selected portions of the NS summaries will be reproduced in the Discussion section below.

TABLE 9.
NATIVE SPEAKER NEWS SUMMARY WORD ANALYSIS

Word or Word phrase	Occurrences	Frequency (%)
people	15	2.3
some, some-	15	2.3
about	9	1.4
know	9	1.4
just	8	1.2
think	8	1.2
country	7	1.1
devastate, -d, -ing	7	1.1
maybe	6	0.9
Poland	6	0.9
communism, -ist, -ists	5	0.8
dead, died, death, dies	5	0.8
happen, -ed	5	0.8

IV. DISCUSSION

a. Teachers' Listening Assessment Practices

Although the sample size for this study was small (11 English listening teachers), the results are significant because there are so few published studies on Taiwan English teachers' assessment practices and this gives us a glimpse into purported listening assessment practices. That being said, some interesting findings can be gleaned from Table 3.

First, checking for main ideas and checking vocabulary comprehension were the most preferred forms of assessment with over 80% of the Teachers indicating they "frequently or exclusively" use these forms. Checking for gist was also popular as 73% of the Teachers indicated they "frequently or exclusively" use this form of assessment. As so many popular listening and speaking books in Taiwan focus on these assessment forms, it is not surprising that teachers would follow this pattern. Assessment types such as these are relatively straightforward and unambiguous and, especially if multiple choice answer formats are used, incorrect answers can be convincingly identified. On the other hand, vocabulary knowledge is important, but it is more critical for reading comprehension than for listening comprehension. For example, quite a number of words are not important for understanding most non-academic oral discourse. As such, the number of vocabulary items for understanding is fewer than many EFL listening teachers might think. In other words, the most important words appear more frequently, but they might appear in synonyms or words with close meanings, i.e. mourn, sorrow, tragedy. Furthermore, recent research has shown that vocabulary instruction is not as helpful as other forms of listening support (Alderson & Banerjee, 2002; Chang, 2007; Chang & Read, 2006; Chen & Tsai, 2012; Lim, 2009).

Second, the least popular forms of listening assessment were checking for accuracy (ie., dictation) and genre identification, with less than half of the Teachers indicating they "frequently or exclusively" use these forms of assessment. As for dictation, just like vocabulary instruction, this result is strange in that it seems to go against much of recent research showing the effectiveness of dictation (Alderson & Banerjee, 2002; Kuo, 2010; Macaro, 2005, p. 178; Prince, 2012). Given that so much work has been done on genre identification and analysis in TESOL, it also seems peculiar that genre identification was lacking in popularity. Since both dictation and genre identification, at least facially, seem just as unambiguous and easy to score as checking for main ideas and checking for vocabulary comprehension, it is unclear why these forms of assessment should be less favored generally speaking besides the possibility that teacher practices have not yet caught up with the literature.

b. Teachers' preferred assessment and analysis of a specific news text

The Teachers were asked to identify which assessment forms they would use for the news story in Table 1. The results of this survey question are not surprising in that main idea check and checking for gist rate so highly among the Teachers. It is surprising that two-thirds of the Teachers would ask for a genre identification given that 72% of the Teachers said they only "sometimes, rarely, or never" use this form of listening assessment (see Table 3). The reason for this is unclear although this specific listening text would seem to be ideal for genre identification as the text itself does not overtly identify itself as a news piece. Additionally, only one-third of respondents indicated they would use vocabulary comprehension or identification assessment for this specific news text. This is curious since more than 80% of the Teachers said they "frequently or exclusively" use vocabulary comprehension or identification to assess listening. This result may be reflective of the limited and repetitive nature of the vocabulary (and synonyms) in the news text.

The Teachers were also asked to select the key main ideas and key vocabulary for the news story. The three most commonly selected main ideas, with more than two-thirds of respondents concurring, included words like death, tragedy, and mourning (Table 5). The main ideas involving the prime minister, archbishop, or neighboring countries were seen by the Teachers as less important. The Teachers demonstrated consistency in their assessment of this news text as the main ideas they selected as important and the main ideas' related words were more or less repeated throughout the reported key vocabulary (Table 6) and the Teachers' model summaries (Table 7). In fact, in the 11 model summaries, there were only five instances of details unrelated to these three key points. The robust nature of this data suggests that the news story, even though it may have involved an unfamiliar subject matter, was unitopical and unambiguous, at least to the Teachers. The affectional and empathic nature of the Teachers' responses indicates that they paid more attention to the emotion transmitted through the story (ie., tragedy, mourning, died, sorrow) than to a deeper subtext within the story concerning Poland's political future (Table 6). The word analysis of the Teachers' summaries (Table 8) reaffirms this point in that death, mourning, tragic, and loss were among the most commonly used words. Overall, the Teachers demonstrated that they can independently produce model summaries with a high degree of similarity in terms of the summaries' main elements.

c. Native speaker news story summaries

Starting with the word analysis of the NS summaries (Table 9), it is immediately clear that the NS did not closely follow the model summaries offered by the Teachers. Specifically, affectional or emotional words such as tragic, mourn, sorrow, and loss are not among the most commonly repeated words (the exceptions are death and devastate). Instead, the NS choose content words of a political nature as the words country, Poland, and communism are more frequently used than "die." This result is also paralleled in the summaries actually made by the NS as only half produced summaries containing the elements "(i) Polish people are in mourning because (ii) a tragedy has struck Poland as (iii) the most important people in the country have died", as was preferred by the Teachers. For example, this NS summary would appear to be in concurrence with the Teachers' as it contains all the necessary elements and no others:

It was like uh tragedy about the death of the two leaders of Poland, how they are kind of coping with their death.

And this summary, although verbose, also has the required elements and little else:

…there was a uh, just something some type of disaster that devastated the people in Poland. People were mourning in Warsaw and they were just like they were just uh upset, grieving, over the loss of a lot of people. A lot of good like maybe politicians I thought, like good people, I guess.

However, only half of the summaries met this standard. The five summaries that differed from the elements contained in the Teachers' model summary focused on what would appear to have been minor details to the Teachers. These five NS transformed the story from one about the Polish people's grief into one about the political future of the country and Poland's relations with its neighbors. This summary was typical:

I would like maybe say that I saw on the news or whatever that there was an attack in Poland of high political- I would think political figures, um because they were following a movement that I caught Russia and then that I know there was another one that they didn't like and they assassinated them to stop the movement and now Poland's mourning and there is concern that they're not going to be able to rise up from this.

The summary below also focused on the elements not highlighted in the Teachers' model summaries:

They are in Poland and something happen where I think they said the leaders of the country I think like 96 leaders of the country had died so they didn't know what was going to happen and like the government, or like, I think they said something about um the European are like uh allies and they didn't know what was going to happen because their government was basically gone.

Since the NS had no problem understanding the text at the word level (as may not always be the case in studies investigating L2 learner output), why does there appear to be such a large divergence in text interpretation between the NS and the Teachers given that the text is facially very direct and unambiguous? This question can be analyzed from at least four possibilities. First, presuming that the NS had no difficulties in bottom-up processing, it is possible that the NS simply did not fully comprehend the news story in the way that the Teachers did, who were under no time constraint. In this case, subsequent additional listening or text reading would produce different results. Without this option, the NS may have resorted to background knowledge to construct their own meaning. Practically speaking, the American education system has little to say on modern Polish history, and usually the only mention of the country comes during study of World War II, the Cold War, and the birth of the Solidarity movement which started in Poland but spread out across Europe and eventually ended in the collapse of the USSR. This perspective of modern Polish history, as a player on the larger European stage, appears to have been what was used by the NS in their interpretation of the news story. Overlooking the Polish people's mourning of the death of their leaders, the NS appear to have resorted to a method of induction, focusing on the political situation and the brief mention in the news story of both Europe and Russia.

The second possibility for the NS divergence was that the NS fully comprehended the news story as presented, but simply interpreted and synthesized it differently than the Teachers. In other words, the NS construction of knowledge was guided by background knowledge different from the Teachers. Since the question was not asked, we do not know the extent to which Teachers considered the alternate interpretation offered by the NS, if they considered it at all. Without arguing the merits of the NS alternate interpretation, it must be taken seriously as the political nature of the event and the notion of conspiracies involving Russia were widely discussed in the media at the time (Thompson, 2010). In fact, as late as October 30, 2012, more than two and a half years after the event and the conclusion of several reports on the crash, it was reported that another group conducting its own independent investigation into the crash had found traces of explosives on the wreckage, potentially implicating Russia and the region's geopolitics (AFP, 2012). It is plausible that the five NS who interpreted the news story to focus on its political aspects had this background information and perspective in mind. Such a different interpretation from the Teachers' is consistent with the notion that integrative and open-ended tasks such as listening to summarize is heavily dependent on individual test-taker characteristics and therefore not objectifiable (Jing, 2010; Yu, 2007).

Third, the Teachers model summary qualitatively differs from five of the NS summaries in degree to which each group repeats chunks of texts and elaborates on the information given. As Rost (1994) and Keck (2006) found, less proficient L2 learners tend to repeat chunks of the target text in their summary rather than paraphrase, interpret, or elaborate. The Teachers' summaries, seen from this perspective, also tended to repeat chunks of the target text more than these five NS summaries. Instead, these five native speaker summaries tended to elaborate and expand on the information given in the news story. A close reading of the news text and comparison to these divergent NS summaries shows that the NS may have over-interpreted what was said in the news text: the news text did not mention a political attack, terrorist incident, or political or governmental turmoil but this was inferred by these NS. This type of attention to detail and elaboration is what Inoue (2009) found among more proficient L2 learners when investigating their ability to orally summarize. It may be that the Teachers' expectations were calibrated for their students' ability level, presumably less proficient and less comparable to native speaker ability.

Finally, Taiwanese learners of English, when engaged in the act of listening, have been found to employ a more empathetic and affectional style (Teng, 2009). In other words, a strategy favored by Taiwanese English learners is to put themselves in the shoes of people involved in the story in order to better comprehend the listening input. Such a tactic may be reflected in the Teachers' emphasis on the emotional qualities of the news broadcast, which is also present in the NS' summaries but not as dominant. The different backgrounds of the study participants may also be reflected in the focus of each group's summaries. The Teachers all share a background in foreign languages and linguistics and the group of natives speakers that produced summaries more focused on affectional aspects were all psychology majors. Of the five NS who attended more to the political aspect of the listening task, only one was a psychology major and the remaining four were studying criminal justice, fashion design, engineering, and journalism, somewhat less emotionally focused subjects. Another possibility along these lines is that the Teachers, with similar educational backgrounds and all being female, represent a segment of the population that is less interested in foreign political matters and therefore did not initially attend to the political aspect of the news story.

d. Implications

The result of the third research question, finding that the NS' summaries diverged between the Teachers' more emotionally-focused model summaries and the political aspects of the news story, is highly significant in that it reminds us of the practical difficulties when using summarization to assess listening. Summarization may be the most practical and useful form of assessment, for reading or writing, as it most closely approximates the demands of real life. However, there can and will be significant divergence among peoples, regardless of language, on what those "critical" elements are. The results of this study show that even within a relatively simple-structured and short news broadcast, there will be numerous interpretations, each with its own epistemological propositions and presumptions. The diversity of interpretations grows exponentially as one moves across populations, through age and culture (Inoue, 2009; Yu, 2007). This study does not begin to scratch the surface of the philosophical issues involved, but it is a warning for teachers that they should tread lightly when attempting to teach and assess synthetic task forms like summarization. In a more internationalized context, recognition of the cultural diversity that characterizes Taiwan university campuses and campuses around the world today is essential to successful English language teaching (see Vorobel & Kim, 2011).

Teaching L2 listening should be different from teaching L2 reading: in real life situations listeners are often not given an opportunity to go back to listen again and again. How we test or assess our students' performance will lead them in the way they attend and focus while they are listening, which, in turn, will form their habit of listening to the foreign language. Therefore, it is more important for EFL teachers to instruct learners how they should listen to different genres of English (eg., which parts of the input requires more of their attention) than to explain to their students what they hear. Hopefully this study has aroused EFL English teachers' attention to the importance of certain specific listening skills training by contrasting the differences between Taiwan EFL teachers' actual listening assessment practice and native speakers' report of their listening performance. Since the results of this study demonstrate the difficulties in reaching a universal consensus on what are the fundamental elements of simple news text, excessive teacher focus on students getting the "correct" answers may ultimately be futile and, even worse, have detrimental effects on student motivation.

It may also be helpful to recognize that, ideally, all alternative forms of text synthesis should be scored equally when the objective is to develop students listening comprehension skills. This is especially important as language pedagogy increasingly turns towards strategy-based instruction and learning, which in turn heavily emphasizes background knowledge and inferencing. If a student relies on flawed (or what is perceived to be flawed) background knowledge in their interpretation of the text and construction of new knowledge, that should not be held against them. When possible, teachers should of course attempt to correct erroneously held beliefs or clarify the shared beliefs of the community which may not be reflected in a particular student's output. However, that is a separate question from whether the student "comprehended" the text. This idealism may seem too much a burden for language teachers and that fact is not to be denied, but if language teachers, especially in Asia and other regions where English is largely a foreign language, are going to break free from the memorize-recite model of shallow language learning, they must consciously heed this phenomenon.

The use of native speakers in this study also presents real challenges that must be confronted. As the notion of the "model native speaker" or "native-like competency" has been eroded in recent years, there may be a desire to dismiss studies such as this. Instead, the use of native speakers reveals that assessing listening ability in all but the most basic and rudimentary fashion will undoubtedly raise serious issues of validity. The native speakers in this study were not used as a normative baseline for assessment purposes, but as a calibration standard against which the measurement tool could be refined. If English learners had been used instead of native speakers, we might have easily dismissed their responses as insufficient, and incorrectly attribute it to the L2 learners' "deficient" listening or speaking ability, such as what happened in Rost's (1994) study. It is highly unlikely that the native speakers in this study did not fully comprehend the words in the news story from a bottom-up processing perspective. Just as so many studies have revealed the problems with using expert native speakers as assessors (Cohen, 1993; Yu, 2007), we have revealed the limitations of what would otherwise appear to be an acceptable, albeit norm-referenced, listening assessment rubric. The native speakers in this study were not called upon for their ability to summarize, but because they would not be encumbered by finite listening and speaking skills, as so many L2 learners are. The native speakers in this study focus our attention on an alternative way to understand the news story and even though we may be uncomfortable with that version, it cannot be dismissed. The tension here is obvious and reflected in Yu's study on democratizing assessment (2007). As language education continues to distance itself from the grammar translation and audio-lingual methods, the profound tension highlighted in this study implores us to return to the assessment drawing board to find a way to better account for these types of discrepancies.

V. Conclusion

This study surveyed 11 practicing Taiwanese English listening and speaker teachers to determine their listening assessment preferences and how they would assess a specific listening task which was a news report. The Teachers showed a high degree of concurrence in their listening assessment preferences and in their assessment of a specific listening task. The sample size was small and the listening task was restricted in length and content, but the Teachers appear to have used an affectional method to comprehend the listening task resulting in model listening summaries that were more oriented towards empathic observations. Additionally, 10 college-aged native speakers were also asked to listen to the same target text and orally produce summaries. Five of the native speaker summaries focused on the same elements as the Teachers' model summaries, and the other five native speaking participants focused more on the political aspects of the listening task's content. This division between native speaking participants correlated with the participants' academic majors.

Acknowledgment

This work was supported by the National Science Council of Taiwan (R.O.C.) [grant NSC-100-2410-H-033-027].

References

[1] AFP. (2012). Explosive traces 'found on crashed Polish presidential jet'. Agence France-Presse.

[2] Alderson, J.C. (2000). Assessing Reading. Cambridge: Cambridge University Press.

[3] Alderson, J.C., & Banerjee, J. (2002). Language testing and assessment (Part 2). *Language Teaching, 35*(2), 79-113.

[4] BBC. (2010, April 10). "Poland in mourning for plane crash victims." Retrieved February 13, 2013, from http://news.bbc.co.uk/2/hi/europe/8613804.stm.

[5] Chang, A.C.S. (2007). The impact of vocabulary preparation on L2 listening comprehension, confidence and strategy use. *System, 35*(4), 534-550.

[6] Chang, A.C.S., & Read, J. (2006). The Effects of Listening Support on the Listening Performance of EFL Learners. *TESOL Quarterly, 40*(2), 375-397. doi: 10.2307/40264527.

[7] Chen, S., & Tsai, Y. (2012). A Country in Focus. *Language Teaching, 45*(2), 180-201.

[8] Cohen, A.D. (1993). The role of instructions in testing summarizing ability. In D. Douglas & C. Chapelle (Eds.), *A New*

Decade of Language Testing Research (pp. 132-159). Washington, DC: TESOL.

[9] Flowerdew, J., & Miller, L. (2005). Second language listening: Theory and practice. Cambridge: Cambridge University Press.

[10] Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: a unit for all reasons. *Applied Linguistics, 21*(3), 354-375. doi: 10.1093/applin/21.3.354

[11] Frost, K., Elder, C., & Wigglesworth, G. (2012). Investigating the validity of an integrated listening-speaking task: A discourse-based analysis of test takers' oral performances. *Language Testing, 29*(3), 345-369.

[12] Inoue, M. (2009). Health Sciences Communication Skills Test: the development of a rating scale. *Melbourne Papers in Language Testing 2009, 14*(1), 55-91.

[13] Jing, Z. (2010). Testing via news videos: an exploratory study. *International Journal of Applied Linguistics, 20*(2), 178-205.

[14] Keck, C. (2006). The use of paraphrase in summary writing: A comparison of L1 and L2 writers. *Journal of Second Language Writing, 15*(4), 261-278.

[15] 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.

[16] Kirkland, M. R., & Saunders, M. A. P. (1991). Maximizing student performance in summary writing: Managing cognitive load. *Tesol Quarterly, 25*(1), 105-121.

[17] Kuo, Y. (2010). Using partial dictation of an English teaching radio program to enhance EFL learners' listening comprehension. *Asian EFL Journal Professional Teaching Articles, 47*, 4-29.

[18] Lim, S. (2009). The Effects of Two Types of Pre-listening Support on EFL Learners" Listening Test Performance-Question Preview and Vocabulary Instruction. *응용언어학 (Applied Linguistics), 25*(3), 365-389.

[19] Macaro, E. (2005). Teaching and learning a second language: A guide to recent research and its applications. New York: Continuum.

[20] Moussu, L., & Llurda, E. (2008). Non-native English-speaking English language teachers: history and research. *Language Teaching, 41*(3), 315-348.

[21] Prince, P. (2012). Writing It Down: Issues Relating to the Use of Restitution Tasks in Listening Comprehension. *TESOL Journal, 3*(1), 65-86. doi: 10.1002/tesj.4

[22] Rost, M. (1994). On-line summaries as representations of lecture understanding. In J. Flowerdew (Ed.), *Academic listening: Research perspectives* (pp. 93-27). New York: Cambridge University Press.

[23] Rost, M. (2002). Teaching and researching listening. Harlow: Allyn & Bacon.

[24] Seidlhofer, B. (1995). Approaches to summarization: Discourse analysis and language education (Vol. 11). Tübingen: G. Narr.

[25] Teng, H. C. (2009). A study of EFL listening styles. *International Journal of Learning*, 16(3), 1-12.

[26] Thompson, D. (2010, April 10). Polish president killed in plane crash: the conspiracy theorists will go crazy. The Telegraph. Retrieved February 13, 2013, from http://blogs.telegraph.co.uk/news/damianthompson/100033738/polish-president-killed-in-air-crash-the-conspiracy-theorists-will-go-crazy/

[27] Van Dijk, T. A. (1988). News as discourse. Mahwah: Lawrence Erlbaum Associates, Inc.

[28] Vongpumivitch, V. (2007). Research on EFL Listening Assessment in Taiwan: Current Issues and Future Directions. *English Teaching & Learning, 31*(3), 63-100.

[29] Vorobel, O., & Kim, D. (2011). Upper-Intermediate-Level ESL Students' Summarizing in English. *TESOL Journal, 2*(3), 330-354.

[30] Yu, G. (2007). Students' voices in the evaluation of their written summaries: Empowerment and democracy for test takers? *Language Testing, 24*(4), 539-572.

[31] Yu, G. (2013). The Use of Summarization Tasks: Some Lexical and Conceptual Analyses. *Language Assessment Quarterly, 10*(1), 96-109.

[32] Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher rates: competing or complementary constructs? *Language Testing, 28*(1), 31-50.

Brent G. Walters holds an M.S. degree in Civil Engineering and a J.D. degree from Ohio State University. He has been teaching at the university level in Taiwan since 2009 and is currently a lecturer at Chung Yuan Christian University where he teaches reading, writing, conversation, and special topics in legal English. His research interests include foreign language listening comprehension, speaking assessment, and care theory.

Ching-ning Chien holds M.A. degrees in Special Education from Tennessee Technological University and in English Education from Ohio State University, as well as a Ph.D. degree in Education from the University of Newcastle upon Tyne, England. She is currently an associate professor in Applied Linguistics at Chung Yuan Christian University where she teaches listening, speaking and reading to freshman students. Her research interests include foreign language listening comprehension, foreign language learning and teaching, bilingualism, phonological awareness and second language acquisition.