Journal of English for Academic Purposes 43 (2020) 100811

Contents lists available at ScienceDirect

Journal of English for Academic Purposes

journal homepage: www.elsevier.com/locate/jeap

Measuring the importance of information in student notes: An initial venture

Joseph Siegel [a, *], Michael J. Crawford [b], Nathan Ducker [c], Naheen Madarbakus-Ring [d], Andrew Lawson [e]

[a] Örebro University, School of Humanities, Education and Social Sciences, 701 82, Örebro, Sweden
[b] Dokkyo University, Japan
[c] Faculty of Humanities, Miyazaki Municipal University, Funatsuka 1 - 1 - 2, Miyazaki City, Miyazaki Prefecture, 880 - 8520, Japan
[d] Victoria University of Wellington, Room 210, Von Zedlitz Building, 26 Kelburn Parade, Kelburn, Wellington, 6012, New Zealand
[e] NIC International College, Japan

ARTICLE INFO

Article history:
Received 17 May 2019
Received in revised form 24 September 2019
Accepted 29 October 2019
Available online 5 November 2019

1. Introduction

English medium instruction (EMI) on university campuses in countries where English is not the first or official language continues to rise. In addition, the opportunity for English as a second language (L2) students to study abroad in English-speaking countries is heavily promoted by both sending and receiving tertiary institutions. Consequently, in both contexts, one of the main jobs of the English for academic purposes (EAP) teacher is to develop L2 students' capacity and tools to succeed in EMI environments. In such contexts, the academic lecture has been and remains the primary method of knowledge transfer (e.g., Bamford, 2005) and the ability to capture and collect information in these lectures via notetaking is indispensable for academic success. As a result, EAP teachers continue to emphasize the skill of lecture notetaking in their courses.

* Corresponding author.
E-mail addresses: joseph.siegel@oru.se (J. Siegel), crawford@dokkyo.ac.jp (M.J. Crawford).

https://doi.org/10.1016/j.jeap.2019.100811
1475-1585/

Notetaking serves two important theoretical functions: the encoding effect that enhances comprehension of input and the storage effect that generates a record of information for later study and use on a variety of tasks (DiVesta & Gray, 1972). The former function operates as the listener takes in information and records it in some way (e.g., verbatim, paraphrasing, illustrations) and occurs nearly simultaneously as they listen. The latter is potentially of more use in the longer term. Once the notes are taken and stored (e.g., in a notebook, folder or digital file), they can be revisited multiple times, revised, reorganized, and used for a variety of institutional and individual tasks over a period of time. Despite the benefits of the encoding and storage effects, notetaking behaviors may be changing due to the widespread availability of technology. For example, lecturers in some contexts may make PowerPoint slides and/or video recordings of lectures available on lesson management platforms (e.g., Blackboard, Moodle); thus students can review the material at their own pace, and in the case of videos, can rewind as often as they would like. However, based on our experiences in teaching and teacher education, it seems notetaking instruction in EAP typically aims at developing students' in-class notetaking abilities.

Notes are first and foremost valuable to the notetaker, yet many EAP teachers may check student notes and make assessments of note quality, either based on intuition or using a priori set of criteria. They may classify notes as "good" for a number of reasons: Notes that are written in clearly legible and attractive handwriting, well organized on the page, make effective use of space and margins, and are potentially useful on upcoming tasks such as summary writing or comprehension questions. Efficient notes, in that they include a maximum amount of information in the fewest possible tokens, may also be considered "good" notes. Conversely, quantity often impacts evaluation as well, as teachers may be pleased to see more (rather than fewer) words on the notetaker's paper. All of these factors, either individually or in combination, might precipitate a teacher to classify notes as "good"; however, there is seldom an objective method applied to the evaluation of content and quality of notes.

A recent paper by Siegel (2018a) reviewed various methods for assessing quality of notes and tentatively promoted the information unit (IU) as the most objective method currently available (briefly, an IU is "smallest unit [of information] one can judge as true or false" (Anderson, 2014, p. 104); a fuller discussion of IUs is available in the Background section). However, the extent to which this method of assessment can be objectively and consistently applied beyond an individual teacher-researcher is in question. Another issue with this method of evaluation of notes is that the EAP teacher is likely not an expert in the field of the lecture and thus may be unable to determine levels of importance within the delivered content. In addition, the notion of an IU seems to imply that all IUs are of equal value within any given lecture, which is unlikely the case.

The present paper aims to build on the IU method of note assessment by applying a three-tiered point system to IUs and investigating the consistency with which a group of international teacher-researchers rated IUs in two TED Talks. This project was undertaken in an effort to determine the extent to which the IU method is viable, valid and reliable, and to move towards establishing a framework upon which note quality can be measured objectively by teachers and/or researchers in the field (i.e., irrespective of individual factors or future tasks on which notes might be used).

The paper begins with an overview of previous research on L2 notetaking, with specific focus on notes-to-task connections, individual notetaker factors, and the need for a system to evaluate note quality. This discussion leads to research questions centered on the methodological viability of the IU as a measure of note quality and describes a study involving a group of five language teacher-researchers who individually rated IUs in two TED Talks using a three-point system. Close analysis of ratings showed considerable variation in how the teachers perceived the importance of individual IUs and suggests that further refinements to a collaborative system of rating is needed. Implications for pedagogy and research are then discussed.

2. Background

Two distinct types of notetaking have been identified: generative and non-generative (Mueller & Oppenheimer, 2014). As the first term suggests, generative notetaking involves creating or generating a reproduction of the information one hears. That is, through paraphrasing and/or summarizing, the notetaker generates their interpretation of the input. Non-generative input requires much less cognitive processing and manipulation of information. Instead, notetakers record verbatim what they hear or see during the lecture. These two types of notetaking may be more or less useful depending on any post-listening tasks: generative notetaking may be preferable for more creative uses while non-generative may support direct recall of information, such as multiple-choice tests.

For the storage effect of the notetaking act to be of use to the notetaker, notes need to be used in connection with some task beyond the notes and/or for review. Teachers may or may not inform their students about the upcoming tasks for which they may require notes. The decision to inform students about a task (or not) prior to notetaking would likely affect the content of the notes. On the other hand, if listeners do not have any particular motivation for taking the notes, they may either record information haphazardly or decide not to take notes at all.

Various types of tasks could be used in connection with the notes a student has taken. One common task would be a multiple choice or gap-fill test of lecture content in order to demonstrate comprehension. If learners knew this type of task would follow a listening and notetaking activity, they would likely try to record as many individual pieces of information as possible. This conjecture is based on the notion that most L2 language students are familiar with multiple choice-type tests and the types of questions and answer choices that often appear on them. Another common post-lecture activity is to write a summary. In this case, specific verbatim factual information is less helpful than broad paraphrasing and truncation of key terms, points, views and/or arguments. Since the purpose is to demonstrate a broad understanding of the lecture, notes would

likely include summaries of each section, and may require reference to extra-linguistic factors such as speaker pitch and tone, hints as to the speaker's attitude towards the topic, and instances of repetition and/or paraphrasing in the lecture. There are numerous other potential post-lecture tasks, including short answer questions, a spoken or written reaction, replicating the lecture and forming counter arguments.

The majority of previous studies on notetaking in an L2 have focused on tasks for which notes were used rather than on the quality of the notes themselves (see Table 1 for an overview of such studies). Moreover, notes are often used in multiple ways by the same student. For example, a student may take notes in order to write a summary based on their notes immediately following a lecture and then use the same notes to review for a summative multiple-choice test at the end of the term. By focusing on the post-notetaking task, these studies emphasize the transfer of information from notes (or the listener's memory) to the various tasks, all of which require slightly different methods of knowledge expression. The present paper diverges from such studies, as its underlying objective is to examine the quality of notes themselves (as opposed to how those notes are used by the note taker) and thus focuses exclusively on note samples and the IUs contained within.

Several previous studies have placed attention on notes themselves (or, in some cases, have dual foci of post-listening tasks as well as notes) (see Table 2).

As demonstrated by the numerous methods for assessing note quality listed in the third column of Table 2, a single objective measure of note quality is debatable. Siegel (2018a) reviewed several of these methods, discussed pros and cons of each, and provided illustrative examples of how the various measures can be applied to the same samples of student notes. These examples show that several of the methods mentioned above (e.g., total notations, total words) fail to capture a meaningful item of information in the way an IU can.

The notion of the IU was introduced by Halliday in the 1960's in reference to functional grammar (Halliday, 2014). For Halliday (2014), the IU, which contributes to understanding systems of information (rather than grammar), is a "separate grammatical unit … [with the] nearest grammatical unit [being] the clause" (p. 115). IU's then are a class made up of multiple words or phrases that carry meaning and may indicate the difference between previously known and new information. In speech, IUs are usually marked with some change in pitch or tone and contain both given and new information (Halliday, 2014).

Siegel (2018a, p. 87), drawing inspiration from Anderson (2014), defines IUs thusly:

"An IU is … defined by Anderson (2014) as 'the smallest unit of knowledge that can stand as a separate assertion … the smallest unit one can judge as true or false' (p. 104). IUs typically contain a combination of at least two words, abbreviations, pictures, and/or symbols, which may include an agent or actor (noun), an action (a verb), and/or a description (an adjective or adverb), the combination of which creates a complete proposition that is explicit and relies more on the written notes themselves rather than on memory to stimulate recall."

However, Siegel's (2018a) assertion that the IU is a preferable method of note analysis would be substantiated through measures of consistency such as inter-rater reliability (e.g., Phakiti, 2015; Révész, 2011) when scoring notes. Building on Siegel's (2018a) application of the IU to notes, and to acknowledge that all IUs are not of equal informational value within a given lecture, the present study adds a three-point rating system for each IU and reports on the ratings made by a group of five teacher-researchers. This study aimed to address the following two research questions (RQs):

RQ1. To what extent do a group of language teacher-researchers interpret the importance of information units (IUs) in lectures consistently?

RQ2. What factors impact language teacher-researchers' decisions regarding the importance of IUs in lectures?

Table 1
Studies that focus on transfer of information from notes to a task.

Study	Purpose	Post-notetaking task
Dunkel, Mishra, and Berliner (1989)	To examine the strength of the encoding effect	30-item multiple choice test
Hayati and Jalilifar (2009)	To examine differences between uninstructed notetakers, Cornell method notetakers, and non-notetakers (see Pauk & Owens, 2013, for a comprehensive explanation of the Cornell method)	TOEFL listening test
Tsai and Wu (2010)	To examine how language (L1 or L2) of notes affects comprehension	Comprehension test
Lui and Hu (2012)	To examine how notetaking affects comprehension test and summary writing performance	Comprehension test; Summary writing task
*Kiewra, Benton, Kim, Risch, and Christensen (1995)	To examine how notetaking format can affect performance testing	Comprehension tests
*Mueller and Oppenheimer (2014)	To compare pen and paper to computerized notes	Factual-recall questions and conceptual application questions
*Bui and McDaniel (2015)	To examine how learning aids (e.g., skeleton outline, illustrative diagram, etc.) affect learning	Free recall test; short answer test

Note: Studies marked with (*) focused on L1 notetaking.

Table 2
Studies that include explicit evaluation of notes.

Study	Purpose	Method for evaluating notes
Dunkel (1988)	To establish and examine differences in L1 and L2 notetaker ability	Total number of words; total number of IUs; test answerability; completeness; efficiency (also included a post-lecture comprehension test)
Clerehan (1995)	To compare L1 and L2 note quality in terms of hierarchical structure	Number of words; hierarchical structure; organization
Song (2012)	To examine effects of different notetaking formats	Note quality compared to a hierarchical system of lecture points (also included a post-lecture comprehension test)
Crawford (2015)	To examine the effects of notetaking practice	Number of content words; notations; abbreviations; highlights; arrows
Siegel (2016)	To examine the effects of systematic teacher-led notetaking instruction using semi-authentic listening materials	Number of IUs
Siegel (2018b)	To examine the effects of systematic teacher-led notetaking instruction using authentic listening materials	Number of IUs (also included a post-lecture comprehension test)

3. Methods

3.1. Participants

The five co-authors of this paper were the participants in the study. Table 3 below provides relevant background information about each participant.

As can be seen in Table 3, the participants are experienced ESL/EFL teachers with teaching experience ranging from 16 to 28 years. In addition, all participants have several years of experience in notetaking instruction. However, three participants have less experience evaluating notes than teaching, and only one participant has used IUs for evaluation.

4. Materials

The materials for the study consisted of two TED Talk transcripts in which all IUs, based on the above definition, were identified by TR1 with highlighting. The first talk, by Prosanta Chakrabarty (2016), was entitled “Clues to prehistoric times found in blind cavefish,” and was divided into a total of 47 IUs (hereafter referred to as Text 1). The second talk, given by Aomawa Shields, was called “How we'll find life on other planets,” and had 46 IUs (hereafter Text 2). The transcripts were generated in Microsoft Word and did not include any information about prosody, pausing, or timing. They only included the words spoken by the respective speaker. Sample IUs from Text 1 included the following: “Ichthyology, the study of fishes”; " I really focused on caves for finding new species”; and “over many, many generations, [cave fish] lose their eyes and their eyesight”. TED Talks were selected because, in our experience, they are popular training materials in university EAP/EFL listening classes and students often self-select TED Talks as listening materials. Still, it is important to note that TED Talks are not typical EMI lectures, and the two genres differ in several ways. The former are usually stand-alone speeches that are well-rehearsed; the latter are situated within a series of lectures, often building on previous sessions and previewing upcoming ones, and, at times, include spontaneous output from the lecturer. Furthermore, one might listen to a TED Talk for enjoyment or general interest, whereas some form of assessment may be linked to an EMI lecture's content.

4.1. Procedures

Participants were e-mailed transcripts for the two TED Talks in which the IUs were highlighted. The two texts chosen were used as materials in a pedagogic project on notetaking (Siegel, 2018b), which involved analysis and scoring of student notes. Therefore, these texts have also been selected for the purpose of this paper, which is to better understand the methodology involved in assessing lecture content and student notes using IUs.

Table 3
Background information about participants.

TR[a]	Nationality/ First language	Current teaching context/country	Years of experience teaching ESL/EFL	Years of experience teaching notetaking	Years of experience evaluating notes	Years of experience using IUs to evaluate notes
TR1	US/English	L2 teacher education/Sweden	18	8	2	2
TR2	US/English	Undergraduate EAP/Japan	28	10	10	0
TR3	UK/English	EMI AND EFL/Japan	18	3	0	0
TR4	UK/English	Pre-sessional EAP/New Zealand	16	7	5	0
TR5	UK/English	Pre-undergraduate EAP/Japan	21	7	7	0

[a] Note: TR = teacher-researcher.

Each IU was preceded by an identification number and followed by a space. For the purposes of this study, participants were instructed to write their evaluations of each IU's importance in the spaces using the scale below. The notion of "importance" can vary depending on the potential post-listening task for which notes will be used. Since no task was explicitly tied to the TED Talks used in this project, the general purpose of the notes was to stimulate recall of content. Participants were asked to judge the importance of content as if they were listeners with the goal of recording and retaining as much information as possible. Since none of the participants were subject specialists in the topics of the TED Talks, they could not determine importance in the same way as an expert in a respective field; instead, their experiences resembled that of students in that they were novel listeners.

3 points = Very important information directly linked to purpose/theme of lecture; without it, some comprehension is lost
2 points = Somewhat important/relevant information
1 point = Slightly important/relevant information that should be in notes, but not crucial if missed

To avoid any inter-rater influence, each participant submitted their scores separately to TR1, who tabulated responses and then shared all collected scores with the group members.

4.2. Data processing

After all of the data was collected, it was then analyzed using statistical and other procedures. To investigate RQ1, first descriptive statistics of the IUs were compiled. This was followed by analyses of consistency in the ratings, first by determining the mode rating for each IU. From there, the number of rating bands from the mode for each IU could be determined. Additionally, Fleiss' kappa was calculated as a measurement of inter-rater reliability.

To investigate RQ2, each teacher-rater was asked to provide an account of their experience using the IU rating scale. Each rater received the following optional prompts and wrote their account:

1. What approach did you use to tackle the task?
2. What did you look for specifically when rating the IUs?
3. What was your experience (pros) of using the IU rating scale?
4. What was your experience (cons) of using the IU rating scale? Was the task difficult? Did you change any of your answers? Did you change your approach between rating Text 1 and Text 2?
5. How confident were you with the ratings?

The qualitative accounts were then analyzed and divided into themes to draw comparisons between each teacher-rater when using the three-point rating system.

5. Results and analysis

5.1. Research question 1

RQ1. To what extent do a group of teacher-researchers interpret the importance of information units (IUs) in lectures consistently?

Table 4 below presents the descriptive statistics of the IU ratings. The "Total" column at the end of each rating for each text was calculated by multiplying the number of "3" ratings by 3, the number of "2" ratings by 2, and the number of "1" ratings by 1. All five TR ratings for each individual IU rating for each text were then added together to provide a total for that rating.

As can be seen in Table 4, the "2" rating is most frequently used to identify IUs in both Text 1 and Text 2. Additionally, all raters then chose the "3" rating followed by the "1" rating for the remaining IUs in each text. This suggests that although each text may have a varying number of IUs, that raters may still identify IUs in the frequency of "2", "3" and then "1" ratings.

However, there is a wide range of variation in the ratings of each IU. For example, for Text 1, the highest number of "3" ratings was 26, and the lowest was half that number at 13. Clearly, while rating independently, in many cases the teachers could not agree on what constitutes a "3" rating. Similarly, for Text 2, the largest number of "3" ratings was 22, and the smallest only 9. Greater consistency was found for the "2" ratings, but not for "1" ratings.

Table 4
Teachers' IU ratings from Text 1 and Text 2.

Ratings	Text 1						Text 2					
	TR1	TR2	TR3	TR4	TR5	Total	TR1	TR2	TR3	TR4	TR5	Total
# of "3" ratings	16	16	13	22	26	93	12	9	11	16	22	70
# of "2" ratings	22	17	27	21	18	105	22	21	16	17	18	94
# of "1" ratings	9	14	7	4	3	37	12	16	19	13	6	66

To allow more detailed examination of consistency in participants' ratings, the mode for each IU was identified, along with the level of deviation across TR ratings. For example, if the mode for an IU was "3," with three people rating it as "3," but Person A rated it as "2" and Person B as "1," then the total for rating bands from the mode would be "3," with Person A being one band from the mode and Person B being two bands from the mode.

Table 5 displays the number of IUs for which 0, 1, 2, and 3 rating bands from the mode were found. Five IUs did not register three identical ratings. For these, the mode is listed as 1.5 or 2.5.

If an IU had zero bands from the mode, it would indicate that all raters had scored the IU the same and the IU was "very consistent", while if a unit was rated as 4 bands from the mode it would indicate that two of the raters had extremely different views about that IU's value. An IU rated thus would be considered "very inconsistent." In the case of Text 1, while 11 of the 47 IUs fell under the category of "consistent" or "very consistent", 29 fell in the grey area of "somewhat" or "neither/nor", perhaps indicating disagreements between the raters were relatively minor. Finally, 7 IUs fell under the rating of "inconsistent", suggesting that these would deserve specific attention if the raters were to further develop this work and try to standardize the IU ratings through a norming activity.

In the case of Text 2, greater consistency was found, with 21 of the 45 IUs found to be "consistent", and only 1 found to be "somewhat inconsistent". While reasons for the difference in the number of "consistent" ratings between Text 1 and Text 2 ratings are unclear, possible explanations could be: (1) raters had unconsciously developed better rating methodology while rating the texts, or (2) that the second text is easier to understand/has clearer delineation between IUs. In the case of the latter, somewhat serendipitously, this may indicate that the raters have found a way of rating the difficulty of academic texts used in EAP courses - the greater the inconsistency, the more difficult the text.

In order to examine the inter-reliability of the ratings statistically, Fleiss' kappa was calculated for both Text 1 and Text 2. For Text 1 the result was 0.05, and for Text 2 it was 0.21. According to the most commonly cited interpretation of results for Fleiss' kappa (Landis & Koch, 1977), the level of agreement for Text 1 is "slight" (0.01–0.20)," and for Text 2 is "fair" (0.21–0.40). Despite the slight increase for Text 2, these figures are clearly lower than an agreeable ideal. Thus, before we can propose IUs as an objective measure of notetaking, it would be necessary to refine how ratings are carried out and to reach a type of collegiate understanding through collaborative dialogue. Potential reasons for the lack of reliability are explored in Research Question 2.

5.2. Research question 2

RQ2 - What factors impact language teacher-researchers' (TR) decisions regarding the importance of IUs in lectures?

The following summaries of the self-reported recalls describe the main findings from the TR accounts.

1. What approach did you use to tackle the task?

The five TRs approached the task similarly to rate the IUs. TR1 thought about how the information relates to the topic, the speaker and the thesis (of the talk) while TR4 focused specifically on identifying the main ideas, key words, and the relevant points needed to understand the talk. In both cases, the TRs looked at how relevant the IU was to the message being conveyed by the speaker. In turn, the stronger the content appeared to be related to the speaker's main objectives, the higher the idea was rated.

The other TRs rated the IUs using systems that reflected writing structures to distinguish similar components presented in the listening text. For example, TR2 and TR3 applied a three-point system to include components similar to a three-paragraph essay, as demonstrated in Table 6.

While it seems that all TRs similarly approached rating an IU based on the academic organization of information, there were important differences that may account for the inconsistent ratings reported earlier. First, as per the approach utilized by TR2 and TR3, an academic essay incorporates a three-tier hierarchy of meaning and focus: topic = 3 pt, explanation = 2 pt, evidence = 1 pt. Although each tier provides greater detail and focus, the unit may add little weight to overall comprehension. While TR1 and TR4 reported that looking for explanations and examples (3 pt) may be crucial to the understanding of the topic, TR2 and TR3 would rate these IUs in the 2 pt and 1 pt categories respectively based on the academic essay approach. Therefore, it is arguable that descriptions of an IU as topic, explanation, or example cannot account for how important an IU is for overall comprehension of the text.

Table 5
Number and percentages of IUs for which bands from the mode were found.

# of bands from the mode	Consistency	# in Text 1 (% of total)	# in Text 2 (% of total)
0	Very Consistent	1 (2%)	4 (9%)
1	Consistent	10 (21%)	17 (38%)
1.5	Somewhat consistent	3 (6%)	6 (13%)
2	Neither nor	24 (51%)	18 (39%)
2.5	Somewhat inconsistent	2 (4%)	1 (2%)
3	Inconsistent	7 (15%)	0
4	Very inconsistent	0	0

Table 6
Connections between IU scoring and essay writing.

TR2/TR3 IU Rating	Three Point Writing System (TR2)	Three-paragraph essay system (TR3)	Example IUs (from Aomawa Shields "How we'll find life on other planets")
3	*The speaker's main idea in introduction*	*What is the topic sentence (equivalent)?*	We don't know what the atmospheres of these planets are like
2	*Explain with information*	*What is the explanation?*	because the planets are so small and dim compared to their stars and so far away from us
1	*Examples, support of main idea*	*What was the actual evidence that supported the idea?*	For example, one of the closest planets that could support surface water – it's called Gliese 667 Cc– such a glamorous name, right, nice phone number for a name – it's 23 light years away. So that's more than 100 trillion miles.

As TR3 mentions, this approach can help listeners work out the topic and then rate information which is interpreted as valuable. However, it is important to recognize that although written and spoken language do not always share the same patterns, this approach could help in identifying IUs. For example, in order to improve the cohesion and coherence of a written text, an academic essay assumes some redundancy of information through the repetition or paraphrase of the topic/key idea while listening to an idea repeated multiple times in a talk may indicate an important point; however, it may not be necessary to note it down each time it is mentioned by the speaker. TR1 noted that if an idea was repeated throughout it would only receive 2 pt, while perhaps under the "academic essay" approach proposed by TR2 and TR3, such an IU would likely receive 3 pt. Therefore, it is open for discussion as to whether repetition could help identify the value of IUs as this may be dilution caused by expected genre conventions.

2. What did you look for specifically when rating the IUs?

All five TRs listed how they defined each of the points of the three-point rating scale, as described in Table 7.

As Table 7 shows, a 3 pt rating was given if it was a unique idea, a key word or supporting idea or directly related to the main idea. A 2 pt rating was awarded if the idea was repeated throughout, was an indirect idea or provided extra/supporting information that could be omitted without affecting the meaning of the main idea. A rating of 1 pt was given if the idea was unrelated or irrelevant to the topic, was personal information related to the speaker or could be omitted without affecting the meaning of the main or supporting ideas.

Similarly, it becomes apparent that three types of methodology were employed by the TRs. For example, in rating an IU as 3 pt, TR4 and TR5 looked at micro-perspective information (such as specific types of syntax, main ideas of talk, objectives/role of speakers, key terminology, statistics, proper nouns, and causal relationships), while TR1 and TR2 looked for ecological perspectives through clear relationships to the topic, and TR3 tried to gain a holistic perspective of how important information was through its relative positioning in the text.

As each TR seemed to look for different things in the IU, clearer directives in deciding how IUs should be identified in academic preparation courses are needed. However, indicating to students how their work will be evaluated, or referring to their own prior experiences of notetaking (learning styles, prior knowledge and personal interest) may impact what is noted during notetaking activities. Thus, in terms of the validity and reliability of IUs as a notetaking evaluation methodology, the question must again be asked "can we objectively grade students' notes?"

3. What was your experience (pros) of using the IU rating scale?

The five TRs also reported on their positive experiences of using the IU rating scale. TR1, who had the most prior experience in developing and using the system in previous projects, observes that it is important to consider how information is presented in lectures as this could impact on how content is understood by listeners. TR2 observes that the process with Text 2 was easier to do as he had experience of this from Text 1. TR3 commented that although the initial rating scale was reasonable, it was important to distinguish whether a piece of information was more important than another, perhaps depending on whether you are listening for gist or listening for details. TR4 expands on TR3's ideas, explaining that although the rating system could be used to define what was important and not when listening to the talk, the ratings could help direct the listener's attention to specific lesson points (e.g. the main idea, specific idea or keyword) to identify what is important.

Two of the TRs in the study, who felt they had familiarity with the texts and IUs (TR1 and TR4), remarked upon the potential for IUs to inform teaching. As teaching activities should be reflected in the assessment methodology and vice versa, their assertion that IUs can help us consider the organization of information and help guide students to finding information indicates the potential of IUs in improving both the teaching and assessment of notetaking. As the analysis in relation to RQ1 suggest, however, further refinement and discussion with peers in the teaching and research field is needed to come to joint agreement on IU rating.

Table 7
TR-reported interpretations of rating scale.

Teacher-Researcher	Rating Scale (3 pt)	Rating Scale (2 pt)	Rating Scale (1 pt)
TR1	Novel idea. Mentioned only once. Related to the main theme	Repeated or related idea throughout	Unrelated idea to the topic or could be unrelated personal info
TR2	Directly related to main idea or supporting idea	Related but perhaps not directly	Connection seemed weak
TR3	Position of information helped judge how important information was: i.e: if IU was at the beginning or end of a paragraph, probably focused on it more (or helped me to focus more)		
TR4	Key words/Key information (proper nouns, names, statistics)	Main idea relevant to 'key' information that would not make sense if omitted	Information that could be omitted and didn't make a difference if included or not
TR5	Main ideas of talk stated, objectives roles of the speakers, key terminology, statistics, proper nouns, causal relationships (clearly signaled)	Supporting details helping to better explain previously mentioned ideas (in form of paraphrasing)	Interesting but not relevant to the main objectives of the talk. (e.g. Supplementary details about speaker's life, comments about names)

4. What was your experience (cons) of using the IU rating scale? Was the task difficult? Did you change any of your answers? Did you change your approach between Text 1 and Text 2?

Four of the five TRs needed multiple reviews of the text in order to rate it as appropriately as possible with TR2 indicating a break of several days, and a whole new approach between the first text and second text. This indicates that a lack of clear distinction between what was very important, important, or somewhat important may have been problematic (see Table 7). Perhaps alternative descriptive systems of rating could be explored in a future study to develop more objective ratings.

Two of the five raters also noted that the process of rating uses a different medium of delivery (i.e., written) than the medium students would receive the text (i.e., aural). This would lead to important differences in both the number of times the text could be revisited (TR3 read the text multiple times) and in the quality of a text (e.g., TR3 and TR4 noted the absence of paralinguistic aspects such as pauses, laughter, changes in pitch, or indeed other visuals that would affect students' ability to focus on key information and discard unnecessary information). Future studies may wish to establish if this has any bearings on the appropriateness, validity, and reliability of the ratings applied.

It was also noted that Text 2 seemed to have more idiomatic, colloquial and anecdotal language so it could be more difficult for students. However, this should not be an issue when examining a text; if we assume each text can stand alone, then rating one text should not have an effect on another text. However, the issue is that the more in dispute the "experts" are about the text, the more we should assume the text is difficult for the students to understand. This claim is potentially supported by the fact that the first text has more inconsistent ratings between each rater than the first text, and thus, could result in more difficulties for students (see Table 4).

At this stage, as an exploratory piece of research investigating the viability of IUs as a valid and reliable way to grade students notes, this finding highlights the difficulty in setting not only appropriate ratings for each IU, but also in the difficulty of each individual rater in setting a consistent approach to rating.

5. How confident were you with the ratings?

TR1 had prior experience of using the IU concept in lectures and with student notes, observing that as with any types of marking or assessment, his confidence and efficiency in using the system improved with frequency. TR4 also had some prior experience of assessing student notes which made some of the ratings easier to apply to the transcript, although suggesting that more specific instructions were (as listed by the TRs in RQ2) could help with consistency and reliability in the ratings.

TR2 was not so confident, describing some of his ratings as just his best guess although he comments that using the three-point approach helped with his ratings. TR3 was also unsure about the reliability of some of his ratings, suggesting that if he looked at his answers again in three to six months, his ratings may not be consistent. TR5 also felt that using pre-identified IUs and his prior experiences in EAP perhaps colored his rating decisions and could influence his decisions to prepare learners for future tests.

6. Discussion

The current lack of objective measures for assessing notetaking quality may reduce students' ability to develop good notetaking skills. Neither clear models of what constitutes good notetaking nor achievement guidelines for notetaking improvement are available. Moreover, pedagogical practitioners currently lack concrete models of what to teach when "teaching" notetaking, and how to do it. Arguably, without objective assessment methodology, students have no way of knowing if they are either (a) doing it right, and (b) improving their notetaking skills. Thus, we argue that clear methodology on how to effectively and objectively grade notes is an essential area for development in EAP and, through this study, have aimed to improve the situation. It has, importantly, revealed inconsistent results and raised questions about how the IU

method may be applied to notetaking in EAP courses and suggests that clear guidelines for determine IUs, applying a three-tiered rating systems, and collaboration among multiple raters are potential steps to improve the viability of the method.

The practice of counting IUs may serve as an indicator of how much of a text a student has understood and been able to take down. However, as a relatively novel way of assessing students' notetaking, various issues related to IUs must be resolved. Firstly, some IUs may hold more importance in a text than others. Given that students are taking notes in a second language, their cognitive resources are likely to be pushed to the limit by the triple focus of listening, encoding, and noting down items. It is, therefore, critical that students are able to focus on the most important IUs when taking notes. As such, this paper proposes a three-tiered ranking for IUs – 3 pt IUs (very important), 2 pt (somewhat important), and 1 pt (slightly important).

In rating Text 1, a wide range of values were attributed to the IUs within the text. For example, from the first text, the highest number of "3" ratings was 26, and the lowest was half that number at 13. In such a case, further work towards determining what constitutes "very important", "important", or "somewhat important" is necessary - without agreement on the value of each IU, the methodology cannot be said to be objectively valid, although it attempts to apply a systematic and justifiable process to the information contained within notes. Similarly, even the task of determining what is an IU can be done more effectively in collaboration.

Reasons for the lack of validity and agreement in rating can be found in the qualitative analysis of the TRs' approaches to rating IUs. Some TRs approached the rating activity with an academic essay in mind, while others focused solely on the inherent value of IUs, while TR5 focused on potential uses of IUs in an EAP course. Further analysis of the TRs' decision making indicated that some had a micro-perspective, recognizing syntax as well as meaning. Others adopted an ecological perspective, considering how the IUs were related to each other. Another TR had a holistic approach to the rating activity, looking for the position and interrelation of IUs within the whole text. At this initial stage, it seems worth recommending that future efforts in this area could perhaps establish from an early stage an agreed procedure or methodology for rating. Once TRs can reach "agreement" on the value of IUs within a text, the IU rating system may be considered more valid.

Being able to consistently rate IUs is also important if IUs are to be used to reliably assess students' notes. In this study, the number of bands from the mode an IU rated was an indicator of how consistent the TRs ratings were for that unit. In the first text, greater inconsistency occurred; 61% of IUs were not clearly rated and 15% were *very inconsistent*. However, in the second text, 47% of IUs were rated *consistently*, a finding suggesting that the reliability of using IUs would improve with practice. In terms of the organization of an EAP course, this indicates, in conjunction with the previous point about validity, that a norming session would be useful for teachers to agree upon an approach, methodology, and specific ratings.

Overall, given that in RQ1, analysis using Fleiss' kappa resulted in inter-rater reliability ratings of slight to fair, the current study did not show that the TRs were able to consistently rate IUs according to this standard measurement. However, the improvement between the first text and second text in both consistency (as measured by the number of bands from the mode) and increases in reliability (as measured by Fleiss' kappa) suggests that a future study with more samples of texts rated by the same group of TRs may indicate that IUs can be reliably used to assess notetaking. A further issue that future research may wish to explore is the question of the "expert opinion". While the TRs in this study have extensive experience in EAP, none of them is an expert in the topics addressed by the texts; thus, future studies utilizing IU ratings by experts in the field (ideally lecturers rating their own content), may help confirm that the TRs' ratings are valid.

Some issues with the use of IUs also arose during the IU rating activities. Firstly, it became apparent that multiple viewings of the text were necessary to rate IUs. This may be due to the fact that the ratings system was relatively novel for the TRs, but may indicate that assigning value to an IU is difficult, and that the use of IUs could potentially be overly sensitive for students who may only listen to a given text once. Secondly, some of the TRs noted that paralinguistic features of a text, as well as prosody and intonation, are often not conveyed by written transcripts. As such, any future study or replication of this work may wish to consider using a form of IU rating based on aural/orally delivered scripts, which could include aspects of Halliday's (2014) tonal aspects of IUs. Thirdly, the notion of "importance" of information is an issue, as this description can mean different things to different people and can depend on assigned tasks.

Several of the TRs mentioned a growing recognition that IUs helped them understand the issues of notetaking, and to better identify key points of the individual texts that they were rating. As assessment and learning are inherently linked, it seems that use of IUs can be developed to both inform teacher practice and help in the development of students' notetaking skills.

7. Conclusion

Few, if any, previous studies on notes (see Table 2) have described and applied the type of systematic multi-tiered analytic procedures described in this paper to IUs in student notes. In other words, most studies simply award one point for an IU without consideration for the relative importance of that item. This study sought to expand on previous work related to assessment of student notes in EAP by investigating the consistency with which IUs in student notes can be rated and exploring the factors that impact a teacher's decisions about IU importance. A group of five EAP teachers rated the IUs identified in transcripts of two TED Talks on a tiered scale representing the importance of each IU. Subsequent statistical analysis demonstrated relatively low levels of consistency, particularly on Text 1. Agreement amongst the group's ratings improved on the second sample, suggesting that consistency of the IU assessment method may improve with experience. Regarding the factors that affect rater choices, each participant responded to a prompt asking about their thought processes

when applying the three-tiered IU system. Findings revealed three distinct approaches (micro, ecological and macro), which provide insights as to TRs thinking about note content, lecture design, information linkage, and EAP course priorities. While this study provides some evidence of the IU system in practice, further thinking at the theoretical level and research at the practical level are needed to advance this area of the academic listening and EAP fields.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jeap.2019.100811.

References

Anderson, J. (2014). *Cognitive psychology and its implications* (8th ed.). New York: Worth.
Bamford, J. (2005). Interactivity in academic lectures: The role of questions and answers. In J. Bamford, & M. Bondi (Eds.), *Dialogue within discourse communities: Metadiscourse perspectives on academic genres* (pp. 123–145). Tubingen, Germany: Max Niemeyer.
Bui, D., & McDaniel, M. (2015). Enhancing learning during lecture note-taking using outlines and illustrative diagrams. *Journal of Applied Research in Memory and Cognition, 4*, 129–135.
Chakrabarty, P. (2016, February). Clues to prehistoric times, found in blind cavefish [Video file]. Retrieved from https://www.ted.com/talks/prosanta_chakrabarty_clues_to_prehistoric_times_found_in_blind_cavefish/transcript?language=en.
Clerehan, R. (1995). Taking it down: Notetaking practices of L1 and L2 students. *English for Specific Purposes, 14*(2), 137–155.
Crawford, M. (2015). A study on note taking in EFL listening instruction. In P. Clements, A. Krause, & H. Brown (Eds.), *JALT2014 conference proceedings* (pp. 416–424). Tokyo: JALT.
DiVesta, F., & Gray, S. (1972). Listening and note taking. *Journal of Educational Psychology, 63*(1), 8–14.
Dunkel, P. (1988). The content of L1 and L2 students' lecture notes and its relation to test performance. *Tesol Quarterly, 22*(2), 68–90.
Dunkel, P., Mishra, S., & Berliner, D. (1989). Effects of notetaking, memory, and language proficiency on lecture learning for native and nonnative speakers of English. *Tesol Quarterly, 23*(3), 543–549.
Halliday, M. A. K. (2014). *Introduction to functional grammar* (4th. ed). New York: Routledge.
Hayati, A., & Jalilifar, A. (2009). The impact of note-taking strategies on listening comprehension of EFL learners. *English Language Teaching, 2*(1), 101–111.
Kiewra, K., Benton, S., Kim, S., Risch, N., & Christensen, M. (1995). Effects of note-taking format and study technique on recall and relational performance. *Contemporary Educational Psychology, 20*, 172–187.
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174.
Lui, B., & Hu, Y. (2012). The effect of note-taking on listening comprehension for lower-intermediate level EFL learners in China. *Chinese Journal of Applied Linguistics, 35*(4), 506–518.
Mueller, P., & Oppenheimer, D. (2014). The pen is mightier than the keyboard: Advantages of longhand over laptop note taking. *Psychological Science, 25*, 1159–1168.
Pauk, W., & Owens, R. (2013). *How to study in college* (11th ed.). Boston: Wadsworth Cengage Learning.
Phakiti, A. (2015). Quantitative research and analysis. In B. Paltridge, & A. Phakiti (Eds.), *Research methods in applied linguistics* (pp. 27–48). London: Bloomsbury.
Révész, A. (2011). Coding second language data validly and reliably. In A. Mackey, & S. Gass (Eds.), *Research methods in second language acquisition* (pp. 203–222). Chichester, UK: Wiley.
Siegel, J. (2016). A pedagogic cycle for EFL note-taking. *English Language Teaching Journal, 70*(3), 275–286.
Siegel, J. (2018a). Did you take "good" notes? On methods for evaluating student notetaking performance. *Journal of English for Academic Purposes, 35*, 85–92.
Siegel, J. (2018b). Teaching lecture notetaking with authentic materials. *ELT Journal, 73*(2), 124–133.
Song, M. (2012). Note-taking quality and performance on an L2 academic listening test. *Language Testing, 29*(1), 67–89.
Tsai, T., & Wu, Y. (2010). Effects of note-taking instruction and note-taking languages on college EFL students' listening comprehension. *New Horizons in Education, 58*(1), 120–132.

Joseph Siegel is senior lecturer in English at Örebro University, Sweden, where he teaches TESOL methodology, linguistics, and applied linguistic research methods courses. He holds a PhD in Applied Linguistics from Aston University. His recent publications have included EFL/ESL listening pedagogy, notetaking, action research, and L2 pragmatics.

Michael J. Crawford teaches in the Interdepartmental English Language Program at Dokkyo University in Saitama, Japan. His primary interest is L2 listening instruction. He is also interested in content-based instruction.

Nathan Ducker is an Assistant Professor at Miyazaki Municipal University where he teaches content classes in intercultural communication and multicultural policy. He is a PhD candidate at Aston University, where he studies willingness to communicate in the Japanese context. He can be contacted at nathanducker@gmail.com

Naheen Madarbakus-Ring is currently a PhD candidate at Victoria University of Wellington. She has taught EFL in Japan, South Korea and the UK. Naheen was shortlisted for the British Council ELTons 2016 Macmillan Talent in New Writing award and received special commendation from the British Council for her previous work researching listening strategies using TED Talks. Naheen was also the recipient of a KOTESOL Research Grant in 2017. Her research areas include focusing on listening in EFL, curriculum and material development, and investigating listening strategies using different texts to create an effective EAP learning environment. Contact Naheen at Naheen.Madarbakus@vuw.ac.nz

Andrew Lawson is an instructor on NIC International College in Japan's intensive one-year EAP program to prepare Japanese students for entry into undergraduate programs around the world. He holds an MA in TESOL from the University of Birmingham and an MA in English and Politics from the University of Glasgow. His research interests include notetaking, the pragmatic expression of disagreement, and vocabulary acquisition. alawson@nicuc.ac.jp