Graded reader comprehension questions and item discrimination analysis

Kris Ramonda and Paul Sevigny

Although the benefits of extensive reading are well known, very little research has investigated the validity of assessment tools to measure general comprehension of graded readers. To address this, quizzes were authored for 42 level 2 Penguin graded readers and administered to 166 students over a semester. Item facility for high-scorers and low-scorers was calculated for the 168 question items from the 42 graded readers, and the resulting item discrimination (ID) score was used to categorize and group quiz items with shared structural and content-based features. The results showed that certain question types tend to be more effective for measuring comprehension of graded readers than others.

Introduction

The confluence of a rapidly expanding role of extensive reading (ER) in language pedagogy with the appearance of new technological platforms has made ER more accessible to L2 learners than ever before. The increasing availability of graded readers through online paperless options is paving the way for ER, whose classroom integration has been considered potentially expensive and cumbersome (see Davis 1995), to be integrated into the classroom far more seamlessly than in previous decades. Not only is ER being brought into the classroom with increasing frequency, but also platforms such as MReader and Xreading provide a means with which to track learner progress, such as how much learners read, what they read, and to what degree they comprehend what they read. These analytics offer valuable diagnostic information to teachers and provide evidence for administrators who are attempting to evaluate and assure the quality of blended learning systems (see Gruba, Cárdenas-Claros, Suvorov, and Rick, 2016).

ER online platforms with integrated learner management systems, such as those described above, hold important pedagogical and administrative implications. In terms of learner behavior and attitude, Brierley (2009) points out that running word counts, or books read, constitute a valid construct in formative assessment for encouraging increased reading. In other words, providing assessments to students for the books they read can potentially lead to positive washback (i.e. increased reading),

ELT Journal Volume 73/3 July 2019; doi:10.1093/elt/ccy062

Advance Access publication February 28, 2019

where washback is the effect that a test has on learning behaviors (see Shohamy, Donitsa-Schmidt, and Ferman, 1996 for further discussion). With regards to course implementation, comprehension quizzes via the aforementioned platforms can be used as feedback to reflect whether the learners understood the gist of the content in the books they read. Such quizzes help teachers to guide a learner's reading level and might also serve as completion checks for books read, and thus count toward some component of a learner's grade in a course.

In light of the potential role quizzes could play in the implementation of ER in a classroom setting, it is important to determine what type of questions tend to be the most effective for measuring reading comprehension of graded readers. Platforms such as MReader provide many thousands of quizzes that have been created through crowdsourcing. Although such crowdsourcing is an effective way of generating a generous pool of question items from which to draw, it also means that the quality of question items can vary from person to person. If individuals do not have question item writing experience or otherwise lack access to guidelines informed by research for that purpose, then it is possible that the question items they produce are not effective measures of general comprehension of graded readers. It has been noted, for example, that some test takers correctly answer multiple-choice comprehension items, ubiquitous in reading comprehension tests, without having read the texts on which they were based (Perkins, 1998). Although a very small number of studies (e.g. Day and Park, 2005) have discussed comprehension questions in an ER context for learner interaction purposes, none have focused on empirically driven data intended to inform the construct validity of comprehension questions as low-stakes completion checks.

One means of addressing the construct validity of comprehension questions in this vein is by exploring the relationship between question type and the validity of these types for measuring the general comprehension of graded readers. This process can be operationalized by carrying out an ID analysis on different question item types.

Question item types

Hillocks and Ludlow (1984) designed and validated a taxonomy of skills involved in the comprehension and interpretation of fiction in an L1 context, which were elicited via various categories of literal and inferential-level question types (reproduced in Table 1). Although initially intended for works of fiction, we maintain that these question types have validity for narrative passages in general (fiction, non-fiction biographies and movie novelizations, etc.) as they set out well-crafted specifications for creating test items that relate item difficulty to specific patterns and features of narrative texts. Moreover, this taxonomy of question types would seem to be well suited and adaptable to ER in light of the fact that the great majority of graded readers follow a narrative style. For these reasons, Hillocks and Ludlow's (1984) taxonomy provides an adaptable framework to measure the relative effectiveness of general comprehension questions in a second-language ER context.

One caveat of adopting Hillock and Ludlow's (1984) taxonomy of skills for the purposes of measuring general comprehension in an L2 context

TABLE 1
Hillocks and Ludlow's (1984) taxonomy of skills in reading fiction

Skill	Brief description
Basic stated information	Textual information without which the story itself could not be possible
Key detail	Textual information signalling significant junctures in the story
Stated relationship	Textual information connecting two characters, two events, or a character and an event
Simple implied relationship	The non-stated relationship connecting two characters, two events, or a character and an event
Complex implied relationship	The non-stated relationship based on inferences from many pieces of information connecting two characters, two events, or a character and an event
Author's generalization	The non-stated implied ideas that connect to the world outside of the book

is that some of the skill-associated question types involve more complex cognitive processes than others and thus might be unsuitable for simply determining whether a student read and understood the gist of a graded reader. In particular, complex implied relationship (CIR) and author's generalization (AG) demand a depth of analysis and understanding that go beyond what might be necessary or desirable. For this reason, the question items in this study focus primarily on the first four categories: basic stated information (BSI), key detail (KD), stated relationship (SR), and simple implied relationship (SIR).

In order to clarify how these four categories were coded, consider an example question item from each category and its relationship to the passages in the story *Apollo 13* (Anastasio, 1999).

Basic Stated Information (BSI)

Conditions stated explicitly and implied many times over.

Item:

What is Jim Lovell's job?

a. Controller
b. Doctor
c. Photographer
d. Astronaut

Text clues (many like the following):

'Yes,' he answered. 'I will bring you a moon rock.'

Jim wasn't the only astronaut on *Apollo 13*.

'In two days that rocket will take you into space.'

Key Detail (KD)

Key details occur at important junctures and causally drive the plot.

Item:

Who was the first person to walk on the moon?

a. Neil Armstrong
b. Fred Haise
c. Ken Mattingly

Text clues:

Photo caption: Neil Armstrong was the first man on the Moon.

Astronaut Neil Armstrong came out of the door and looked up into black space. He looked down at the gray rocks … He was the first man on the moon.

Stated Relationship (SR)

A relationship between two story elements that is only stated once, but directly.

Item:

Who said this: 'Houston, we have a problem.'

a. Jim Lovell
b. Jack Swigert
c. Fred Haise

Text clues:

Jack spoke into the radio. 'Houston, we have a problem.'

'Say again,' Mission Control answered.

'Houston, we have a problem.'

Simple Implied Relationship (SIR)

Same as SR but stated relationships and causes must be inferred.

Item:

Jim Lovell brought a moon rock back to earth.

True
False

Text clues:

'They're not going to the moon,' said the Controller.

'We're not going to the moon,' he said.

'My father's coming home. My moon rock isn't important. He's coming home!'

Item discrimination analysis

One means of determining how well an assessment item is working is by ID analysis. The rationale behind ID is that if a particular assessment item can discriminate well between high and low achievers, with high achievers answering the item correctly and low achievers answering it incorrectly, then the item is working as intended. Stated differently, those students who consistently answer assessment items correctly overall on a given assessment are expected to be indicative of high achievers. Conversely, those students who repeatedly answer incorrectly throughout an assessment could be thought of as low achievers. The underlying assumption is that high achievers are on the whole more likely to genuinely comprehend and successfully answer any one assessment item based on their cumulative record of success on the entire assessment. Since low achievers exhibit the opposite trend of consistently answering incorrectly on the whole, they are more likely to answer any one item incorrectly. It is important to note that low achiever does not necessarily reflect aptitude, as an overall low score could simply be the result of not preparing adequately.

The range of values for ID is from −1 to 1, in which 1 is a case where all the high achievers answer a given item correctly, and all the low achievers answer it incorrectly. A value of −1 is the reverse of the above and is indicative of an extreme case in which the item is not working at all as intended. The cutoff points for determining the high- and low-achiever groups can vary, but the top and bottom quartile range of scores on a given assessment is often used (Brown, 2005), and that is how these cutoffs were operationalized in this study. Furthermore, as each individual quiz only had four questions, the single assessment from which the top and bottom quartile ranges came included the question items for *all quizzes combined*. This is because dozens of observations provide a more reliable picture of which participants were high and low achievers.

In sum, Hillocks and Ludlow's (1984) taxonomy of skills offers a means to operationalize the construct of question type, and ID score is a reliable tool with which to measure question type effectiveness. With this in mind, the following research questions have been posed:

1 *Which question type tends to produce the highest ID scores for the general comprehension of graded readers?*
2 *What possible factors might influence the ID scores for different question types?*

Participants

One hundred sixty-six Japanese first-year university students took part in the study. These students were streamed into low–intermediate level classes based on TOEFL scores taken at the start of the semester.

Procedure

Graded readers were implemented across several sections of low–intermediate reading-focused classes. A total of 168 physical copies (four copies of 42 titles) of Penguin level 2 were acquired and made available to the teachers in charge of the sections involved in this study. Each week of the first six weeks of the semester, students selected a book to read from the 42 available titles. Students were allowed time at the end of class to begin reading the book they chose, but most of the reading took place outside of class.

In order to determine whether students completed the books and understood the content, the authors and one other teacher created an initial bank of 168 quiz items based on the 42 titles of graded readers used in this study (four question items per reader). Since the aim of the quizzes was to check for general comprehension, it turned out that only the first four of Hillocks and Ludlow's (1984) taxonomy of question types (BSI, KD, SR, and SIR) were reflected in the question items. In terms of format, all question items were expressed either as multiple-choice or true/false. Once all the question items were written, the quiz authors met to discuss the pool of items from all the quizzes and eliminate any potentially problematic ones, such as items for which the answer could possibly be given away by pictorial clues on the front cover or by the summary provided on the back cover of the book.

After the six-week treatment concluded, the authors compiled the quiz scores and carried out an ID analysis. Quiz items were grouped into one of four categories based on their ID. Specifically, those quiz items with an ID of –1.0 to 0.0, 0.1 to 0.24, 0.25 to 0.49, and 0.50 to 1.0 were placed into the categories of very low, low, acceptable, and good item discriminators, respectively. The cutoff points for these ranges were chosen because this would allow for a relatively equal distribution of question items into each category for comparison. Once the quiz items were grouped, two independent coders (the authors of this study) proceeded to code each question item based on Hillocks and Ludlow's (1984) question types. Following this, the coders met to discuss and resolve any discrepancies in coding.

Results

Table 2 and Figure 1 present the data for the ID categories along with the associated question types comprising each. As the number of questions types in each category is unequal, the relative gains in terms of percentages are more informative than are the raw number of items for each range of ID. For this reason, the percentages of individual question type by category are indicated in both Figure 1 and in parentheses next to the number of items for each ID range and for the total in Table 2. The percentages across ID ranges appear to be fairly stable for both KD and SIR, which is unsurprising. However, there is a noticeable contrast in the proportion of BSI and SR question types. Although SR only comprised 19.5 per cent of the total questions, they made up nearly a third of acceptable and good question item types. Conversely, the opposite trend appears to be the case for BSI, in which the question type accounted for 29.8 per cent of the total questions but represents 40.5 and 37.5 per cent of the very low and low question types, respectively.

The trajectory of these trends can be visually observed in Figure 1. Although KD and SIR do not appear to diverge much, there is a substantial drop and subsequent rise in BSI, resulting in a non-linear trajectory, which will be further addressed in the discussion. Of most interest is SR, which shows a steady rise up to acceptable, where it flattens out. This shows that, in terms of proportion, the SR question type yielded comparatively effective quiz questions.

Discussion

From the results, it can be seen that the overwhelming majority of very low and low ID question types were BSI and KD, cumulatively. However, this is unsurprising when considering that these two question types together comprised 71.5 per cent of the total pool of questions. Furthermore, although SIR question types appeared only a few times in the very low and low ID categories, this is also expected as there were only

TABLE 2
Distribution of question type by ID score

Category	Very low	Low	Acceptable	Good	Total
ID range	–1.0 to 0.0	0.1 to 0.24	0.25 to 0.49	0.5 to 1.0	
N	37	48	40	43	168
BSI	15 (40.5%)	18 (37.5%)	5 (12.5%)	12 (27.9%)	50 (29.8%)
KD	16 (43.2%)	19 (39.6%)	19 (47.5%)	16 (37.2%)	70 (41.7%)
SR	3 (8.1%)	6 (12.5%)	12 (30%)	12 (27.9%)	33 (19.6%)
SIR	3 (8.1%)	5 (10.4%)	4 (10%)	3 (7.0%)	15 (8.9%)

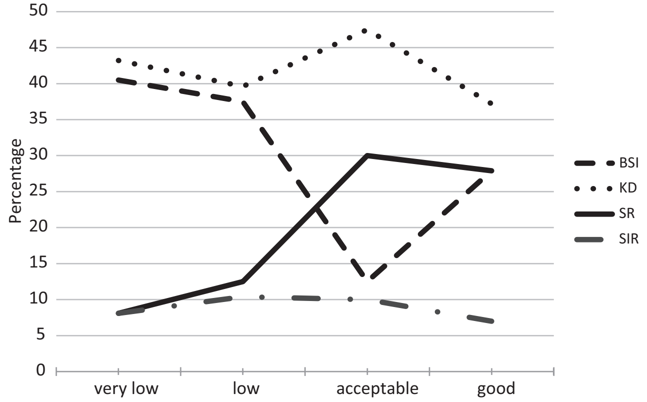

FIGURE 1
Graphed distribution of question type by ID score.

15 question items that were deemed to be SIR. In fact, it would be difficult to draw any firm conclusions regarding SIR, given the very small sample size included in this study (see limitations for further discussion). On the other hand, SR appears to show a very encouraging trend in that, despite having few instances of appearing in the very low and low ID ranges, SR occurs 12 times in both the acceptable and good ID ranges. Thus, while BSI, KD, and SIR question types have very mixed results, the SR question type appears to relatively measure comprehension of graded readers effectively, at least within the confines of this study. But what does this mean and how can this be accounted for?

In order to interpret what these results mean and how they can contribute to answering the second research question posed in this study, it is important to draw attention to two types of cases that can cause the ID for a question item to be low (~0.0). As an extreme case, consider an item that is easy to the extent that all students get the item correct. It could be the case that all the students read and understood the book for which the item was created. Although this is certainly feasible, it is more likely that the item itself was able to be answered correctly without having read the book, making it ineffectual for measuring comprehension. In such an extreme case, the ID would be 0.0 since the correct answers by the high achievers would be precisely offset by the correct answers by the low achievers. Conversely, the other case which could produce an ID of 0.0 is when an item is so difficult that all the students answer the item incorrectly. This case would almost certainly be undesirable as not only are the high achievers expected to be able to answer an effective item correctly, but the resulting low scores might further have a demoralizing effect on those who did actually read and understand the book in question. As for ID scores of below 0.0 (very low), it goes without saying that these should be discarded. In this study, such undesirable ID scores can be accounted for in part by honing in on potential question types that run a greater risk of being too easy or too difficult for participants to answer.

It could be, for example, that skill-associated question types such as BSI and KD are more susceptible to information revealed through pictorial or summarizing information on the cover or in the book that obviate the need to read the story in order to successfully answer the question.

Considering the centrality of information that is considered 'basic' and 'key' to the story could mean that such information can be surmised merely by scanning the text. This is related to what Bell (2011) refers to as a 'preview level' understanding, which could prematurely reveal answers to BSI question types through a cursory examination of headings, forewords, back-cover blurbs, chapter titles, and illustrations. Although the authors attempted to curtail the number of such question items by reading each book they authored questions for, it is entirely possible that subtle cues that could give away answers to questions of this type were overlooked. Conversely, given that SR is neither considered 'basic' nor 'key', this type of information might be less obvious or less likely to be transmitted through such preview-level information, rendering it less likely to be answered correctly by someone who did not successfully read and complete the book in question.

Furthermore, we have identified a number of SR question item subcategories, which item writers might find useful. The examples below were all in the acceptable or good ID range with the scores reported to the right of each question stem. Statements were true/false question items.

(*Character and Statement)*
Who said, 'Houston, we have a problem'? (0.75)
(Character and Motivation)
Who does Algernon later want to marry? (0.6)
Anne wants to marry Mr Elliot. (0.6)
Who does Captain Wentworth love? (0.4)
Who did Thumbelina marry? (0.4)
Why were the birds waiting at Nat's window? (0.6)
(Character and Relationship)
What is the relationship between Jonathan and Evelyn? (0.4)
Sherlock Holmes is the storyteller in these stories. (0.8)
(Action and Location)
Where did the policemen take the animals? (0.5)
(Time and Action)
The birds attack during the day, but not at night. (0.4)
(Character and Status)
The Railway Children's father is a spy. (0.4)

Although this study used a multiple-choice and true/false question format, a potential alternative means of organizing SR questions for a comprehension quiz could be through using a matching question such as the following from *Apollo 13* (Anastasio, 1999):

Match the astronauts and their actions or conditions in the story:

Neil Armstrong Fred Haise Jim Lovell
Ken Mattingly Alan Shepard Jack Swigert

a. ___ He was the first human on the moon.
b. ___ He has a problem with one of his ears.
c. ___ He takes Alan Shepard's place.
d. ___ He is Apollo 13's lunar module pilot.
e. ___ He can't fly on Apollo 13 because he never had the measles.
f. ___ He is the new command module pilot.

To summarize, categories of distractors for SR matching questions could include character names, times of events, locations, types of relationships, actions, and quotations.

Limitations

This study has some limitations that should be addressed in order to help direct future research in this area. One such limitation is the process by which this study was conducted. Given the large-scale, course-wide implementation of this research, our primary concern was for the betterment of the EFL program at the institution where this study was carried out. Thus, when we first created the bank of questions for the quizzes, our primary concern was to create items based on (1) our experience writing quizzes and (2) the type of questions the contents of each book seemed to lend itself to. In other words, we produced question items based on our teacher intuition and without a strict regard for a precise balancing of question type or question format. The result of this is that we ended up with an unequal number of total question types which makes the overall design less controlled and more subject to error. Future research might aim for a smaller, more controlled version of this study to see if similar results can be replicated.

Secondly, upon coding the question items according to Hillocks and Ludlow's (1984) taxonomy of skills, it became readily apparent that not all items fit neatly into a particular category. Although follow-up rater discussions perhaps helped to mitigate any errors attributable to this, there were a few cases in which we had difficulty arriving at and deciding on a corresponding category. In light of this it is advisable either to increase the number of rater judges or perhaps even modify Hillocks and Ludlow's (1984) taxonomy of skills in order for it to be more inclusive for the purposes of measuring the general comprehension of graded readers.

Conclusions

The increasing availability of graded readers through virtual libraries (e.g. Xreading) and the ease of assessment through graded reader quiz sites (e.g. MReader) have made it easier than ever before to integrate extensive reading into the classroom. Although general comprehension quizzes can act as a check that students have read and understood a graded reader, it is unclear whether certain question types are more effective than others.

This study examined the effectiveness of different question types for measuring the general comprehension of graded readers. The findings here suggest that while all four question types (BSI, KD, SR, SIR) are viable, SR was the most effective at eliciting high ID scores. One contributing factor for this could be that the SR question type is less susceptible to having its correct answer given away through pictorial or summarizing information in a preview-level cursory viewing. Alternatively, it might simply be that the SR question types are easier to author with plausible distractors. Further research is needed that draws on qualitative data in order to shed light on what factors might hold the most influence for the relative effectiveness of question type. In addition, subtypes of SR question types as identified in this study (character and motivation, action and location, etc.) might be investigated to determine whether certain ones tend to elicit stronger ID scores than others. Such avenues of inquiry could provide for a more fine-grained understanding of the relationship between question type and item validity in an ER context, as well as offer teachers practical guidelines when creating general comprehension question items for graded readers.

Final version received October 2018

References

Anastasio, D. 1999. *Apollo 13*. London: Penguin.
Bell, A. 2011. 'Re-constructing Babel: discourse analysis, hermeneutics and the interpretive arc'. *Discourse Studies* 13/5: 519–68.
Brierley, M. 2009. 'Assessing extensive reading through written responses and comprehension tests' in E. Skier and T. Newfields (eds.). *Infinite Possibilities–Expanding Limited Opportunities in Language Education: Proceedings of the 8th Annual JALT Pan-SIG Conference* (pp. 45–53). Chiba, Japan: JALT PanSIG.
Brown, J. D. 2005. *Testing in Language Programs: A Comprehensive Guide to English Language Assessment*. McGraw-Hill College.
Davis, C. 1995. 'Extensive reading: an expensive extravagance?'. *ELT Journal* 49/4: 329–36.
Day, R. R. and **J.-S. Park**. 2005. 'Developing reading comprehension questions'. *Reading in a Foreign Language* 17/1: 60–73.
Gruba, P., **M. Cárdenas-Claros**, **R. Suvorov**, and **K. Rick**. 2016. *Blended Language Program Evaluation*. Basingstoke: Palgrave Macmillan.
Hillocks, G. and **L. H. Ludlow**. 1984. 'A taxonomy of skills in reading and interpreting fiction'. *American Educational Research Journal* 21/1: 7–24.
Perkins, K. 1998. 'Assessing reading'. *Annual Review of Applied Linguistics* 18: 208–18.
Shohamy, E., **S. Donitsa-Schmidt**, and **I. Ferman**. 1996. 'Test impact revisited: Washback effect over time'. *Language Testing* 13/3: 298–317.

The authors

Kris Ramonda is an associate professor of applied linguistics at Kansai University in Japan. He completed his PhD at the University of Birmingham where he investigated the relationship between picture-based input and the learning of figurative expressions. His research interests include vocabulary acquisition, metaphor, and extensive reading.
Email: ramonda@kansai-u.ac.jp

Paul Sevigny is a tenured lecturer, coordinator, and language researcher at Ritsumeikan Asia Pacific University in Japan. He has an MA from the University of Hawai'i and is completing a PhD at the University of Birmingham. His research interests include discourse analysis, stylistics, extensive reading, and vocabulary.
Email: paul.sevigny@outlook.com