Reliability and Validity of "Invisible-Gap Filling" Items

Shizuka Tetsuhito
Kansai University

ABSTRACT

The purpose of this study was to explore the potential of "invisible-gap filling" items primarily as an in-house achievement measure of reading-oriented courses and secondarily as a more general overall-ability measure. More specifically, it compared multiple-matching "invisible-gap filling" items and their "visible" counterparts in terms of item facility, item discrimination, test reliability, and test validity.

Eighty-eight Japanese university 1st year students took a 25-item invisible-gap filling test and its visible counterpart, along with two 25-item c-tests, the combination of which constituted a semester-end examination of a reading-oriented course. The invisible and visible gap filling tests were based on the same passage covered in the course. Target words (i.e., words to fill the gaps) were also the same between the versions, making the salience of the gaps the only difference between the two. Hence, psychometric property differences between these two versions, if any, should be attributed to the gap visibility condition difference. One c-test was created from a passage already covered in class and the other from a new passage. The former served as an achievement criterion while the latter was considered a proficiency criterion.

Results indicated that the invisible-gap filling items had (1) lower facility values, (2) higher discriminations, (3) higher reliability, (4) higher validity as an achievement measure, and (5) higher validity as a proficiency measure, than its visible counterpart. Based on these findings, it is contended that invisible gap filling is a technique that can be used to produce reliable and valid achievement tests with relative ease. After discussing possible limitations of the format, two possible modifications are proposed.

1. INTRODUCTION

1.1. Requisites of EFL Reading Achievement Tests

Achievement tests are tests "directly related to language courses, their purpose being to establish how successful individual students, groups of students, or the courses themselves have been in achieving objectives" (Hughes 2003: 13). Falling within this definition are semester-end tests of reading-oriented courses at middle schools and universities in Japan. In such tests, a common practice is to present parts of reading texts already covered in class and require students to answer miscellaneous questions related to them, involving pronunciation knowledge, word meaning, grammatical structure analysis, local comprehension, etc.

Although some discourage using materials already covered in class for achievement tests (e.g., Hughes 2003), the practice seems necessary, or even desirable, at least in some EFL contexts. If a final achievement test of a reading course uses no part of the passages read in class, it belies what Japanese students expect of a course-end test and could be interpreted as a statement that those passages are not worth reviewing before the exam. That definitely is a message that we want least to convey to our students, since we want them to read those used materials over and over again even after the class, *with* comprehension rather than *for* comprehension. When so doing, we want them to confirm that they can accurately and efficiently process their meanings, based on accurate and speedy parsing and smooth meaning retrieval of lexical items. Such repeated reading is reported to raise L1 (Perfetti 1985; Meyer and Felton 1999; Samuels 2002) and L2 (Taguchi 1997) reading fluency.

However, exactly because they are already dealt with in class, such used materials do not lend themselves readily to ordinary content-focused items. Even when such items are responded to correctly, it is not certain whether the successful responses are based on accurate reading on the spot or from memorizing the content itself, already revealed and shared in one way or another in class. Even when students satisfactorily translate part of the passages into their L1, we cannot be sure if they actually have processed the target part in L2 or they are just writing from memory of the content itself.

Hence, in an EFL achievement test context, a test technique that meets the following requirements would be an asset. First, even when based on covered materials, the technique produces items that cannot be answered from memory of the content alone. Second, it creates a positive backwash of encouraging learners to pay attention to the target language itself, rather than to its L1 translation. In addition to these two essential conditions, adding to its practical value would be relative ease with which to create items. One technique that seems to meet all these requirements is the "invisible-gap filling" proposed by Shizuka (2002: 205-206).

1.2. Invisible-Gap Filling Technique

Invisible-gap filling technique, as opposed to ordinary (or, in this paper, *visible*) gap filling, is defined as a testing method that requires candidates to fill in grammatical or semantic "gaps" that are not indicated by brackets, underlines, or any physically noticeable symbols. Visible and invisible gaps can be exemplified as follows:

A sentence with a visible gap

> *The purpose of the research was to (   ) a new item type.*

The same sentence with the corresponding invisible gap

> *The purpose of the research was to a new item type.*

Obviously, the existence of a visible gap can be perceived by anyone with ordinary eyesight, even by those who do not know a single word in the language, whereas to detect the location of an invisible gap, reading ability high enough to comprehend the text at least to some degree is required, let alone to fill it in with some appropriate word.

One way of using this technique for creating reading test items is to remove an appropriate number of words--one word from one position--from a passage thereby creating so many invisible gaps, and place those words below the text in a random order. An example follows:

> If there is a picture of you somewhere on the Web, someone with common image-editing software, and a little time and expertise, can easily paste your onto someone else's body and repost the image.
>
> "Faked" nude photos are more commonly associated with supermodels and celebrities, everyday individuals can also fall victim, as Mark Hall, a warehouse manager in Connecticut, and his wife Nancy learned.
>
> Hall, 33, and his wife put a picture of her on Rankpeople.com, a Web site that allows to look at pictures of people and rate their appearance a one-to-10 scale.
>
> Within a day or two, Nancy called her husband at work to tell him someone had altered the image so that appeared naked, reposted it on the site, and anonymously e-mailed her Rankpeople.com account.

Removed words: on; head; she; users; but

The test-taker's task is to answer the position each word was removed from. If order to successfully respond to these items, they have first to identify the invisible gaps, and then to select the appropriate word to fill each of them. There is no question that generally these items have more to do with careful reading than with "expeditious reading" (Hughes 2003: 138), but at a closer look, these items seem to tap different aspects of reading depending on the test-writer's choice of the invisible-gap locations. Gaps can be created to test fairly elementary word-level comprehension;

. . . can easily paste your onto someone else's body and repost the image. . .

(Key: *head* should be placed after *your*)

to test knowledge of word usage;

. . . Rankpeople.com, a Web site that allows to look at pictures of people . . .

(Key: *users* should be inserted after *allows*)

or to tap correct understanding of idea linking that involves a somewhat larger context.

. . . "Faked" nude photos are more commonly associated with supermodels and celebrities, everyday individuals can also fall victim, . . .

(Key: *but* should be placed before *everyday*)

It may be worthwhile to compare the above five-item invisible-gap test with its visible counterpart, which is presented below. First, the visible version is predicted to be generally easier because the gap locations are already given. The task left for the test-taker is only to find an appropriate match from among the five options. Second, the visible version does not necessarily require the candidate to process everything in the text. It is well possible to look only at immediate context of each gap and select the correct word. In the invisible version, on the other hand, every word needs to be processed in order to find where the gaps are located in the first place. Finally, unlike the visible version for which random guessing will on an average result in 20% correct, the invisible version can practically be considered guessing-proof, with the number of possible gap locations being as many as the number of all the words in the passage plus one.

> If there is a picture of you somewhere on the Web, someone with common image-editing software, and a little time and expertise, can easily paste your (   ) onto someone else's body and repost the image.
>
> "Faked" nude photos are more commonly associated with supermodels and celebrities, (   ) everyday individuals can also fall victim, as Mark Hall, a warehouse manager in Connecticut, and his wife Nancy learned.
>
> Hall, 33, and his wife put a picture of her on Rankpeople.com, a Web site that allows (   ) to look at pictures of people and rate their appearance (   ) a one-to-10 scale.
>
> Within a day or two, Nancy called her husband at work to tell him someone had altered the image so that (   ) appeared naked, reposted it on the site, and anonymously e-mailed her Rankpeople.com account.

Removed words: on; head; she; users; but

1.3. Research Questions

The present study was designed to empirically examine psychometric properties of invisible-gap filling tests as an easy-to-create achievement measure in a Japanese EFL context. It attempted to shed light on this type of test by contrasting it with its visible counterpart or multiple-matching rational cloze procedure. More specifically, it addressed the following research question: How does the gap visibility condition difference affect the gap-filling items in terms of (1) mean item facility value, (2) mean item discrimination, (3) test reliability, (4) validity as an achievement measure, and (5) validity as a proficiency measure?

2. METHOD

2.1. Subjects

Eighty-eight Japanese university 1st-year students enrolled in English II taught by the author served as subjects as a class requirement. The primary focus of the course was on improving their reading ability.

2.2. Instruments

A 25-item invisible-gap test, a 25-item visible-gap test, and two c-tests consisting of 25 items each, were prepared. All the passages were taken from the textbook for the course, *Reading Communicator* (Shizuka 2002b). The passages and the removed words (i.e., the target words) for the invisible and the visible tests were identical. The passage used for the invisible- and visible-gap filling had already been covered in class. Of the two c-test passages, one had been covered but the other had not. The c-test based on the "covered" passage served as an achievement criterion, whereas that based on the new passage served as a proficiency criterion. That is, the first c-test was expected to reflect the general degree to which the students had achieved the course objective, by fully processing the target materials and learning, where appropriate, new grammatical structures and vocabulary items, and the second c-test was meant to predict their ability to process unfamiliar materials of comparable readability. All the tests are found in the Appendix.

2.3. Procedure

The whole set of tests was administered in one sitting (60 min.) as the semester-end examination of the course. The invisible- and visible-gap filling tests were printed on the same sheet of paper, with the invisible version on one side and the visible version on the other. Examinees were instructed to tackle the invisible side only during the first 30 minutes. When the 30 minutes passed, they were allowed to turn over the sheet to begin the visible side. It was emphasized that each side should be responded to strictly in the designated slot alone and any non-compliance with the instruction -- e.g., looking at side 2 in period 1 or going back to side 1 in period 2 -- would be considered cheating. The c-tests were distributed approximately at minute 45. The instruments and the procedure are summarized in Table 1.

Table 1. Instruments and data collection procedure

Time		k	Material familiarity	Length
30 min.	Invisible-gap filling test	25	familiar	453 words
30 min.	Visible-gap filling test	25		
	Achievement C-test	25	familiar	77 words
	Proficiency C-test	25	unfamiliar	95 words

k: the number of items

2.4. Scoring and Analysis

Invisible-gap filling items were initially planned to be scored on a partial credit basis. If a gap was correctly located, 1 point was to be awarded; if a filler was correctly chosen in addition, another 1 point was to be given resulting in a full credit of 2 points for the item; if the gap was not located correctly, no point was to be given. However, the actual marking of the test revealed negligibly few cases eligible for the partial credit. It was virtually always the case that a correctly located gap also meant a correctly filled one. Hence, deemed not worthwhile, partial credit scoring was dismissed and simple dichotomous scoring was employed instead. For the other four tests, dichotomous scoring was the only option in the first place.

Analyzed were (1) the relationship between the performances on the invisible and the visible tests, (2) the degree to which performance on the invisible test predicted performance on the achievement c-test and the proficiency c-test, respectively, and (3) the degree to which the performance on the visible test predicted performance on the achievement c-test and the proficiency c-test, respectively.

3. RESULTS

To our regret, one item in the visible gap test was found faulty (see Appendix) after the test administration and hence had to be dropped from the subsequent analyses, together with the corresponding invisible item. Therefore, the number of items in both tests was reduced to 24.

3.1. Item Facility

First, item facility values were examined. It was expected that invisible gaps would make harder items, which in fact turned out to be the case. The descriptive statistics are shown in Table 2. For the visible items, the values ranged from .522 to 1.000, with the mean at .891. For their invisible counterparts, the range was much larger with the minimum at .068 and the maximum at .875 and the mean was .491. The standard deviations were .145 for the visible items and .223 for the invisible items. A t-test revealed that the difference between the means was highly significant (t = -9.57, p = .000, two-tailed).

Table 2. Descriptive statistics of item facility values

	k	N	Mean	SD	Max	Min
Visible	24	88	0.891	0.145	1.000	0.523
Invisible	24	88	0.491	0.223	0.875	0.068

k: the number of items

To get a better picture of the visible-invisible relationship, the two facility values relating to each gap were graphically represented (Figure 1). First, the facility values were higher when the gap was visible for all but one case. Of passing interest here is gap 14, for which the values were identical at .523. This means that the additional burden of locating the gap itself did not contribute to making the task any more difficult. The gap in question was between *every* and *to* in the following context.

> *"Rankpeople takes every to make sure that nothing up there is . . .*

Probably that *every* should be followed by some noun was too obvious for the subjects in this study. It is interesting to note that this gap made the most difficult visible item and an invisible item of only mediocre difficulty. Second, the ceiling effect is clearly observable for the visible items, with as many as 17 out of 24 items having a facility value of .9 or higher. Four of them were successfully responded to by all the candidates. Third, one may sense that despite the ceiling effect for the visible items there appears to be a weak positive correlation between the two variables. In fact, the Pearson correlation was significant (r = .445, p < .05).

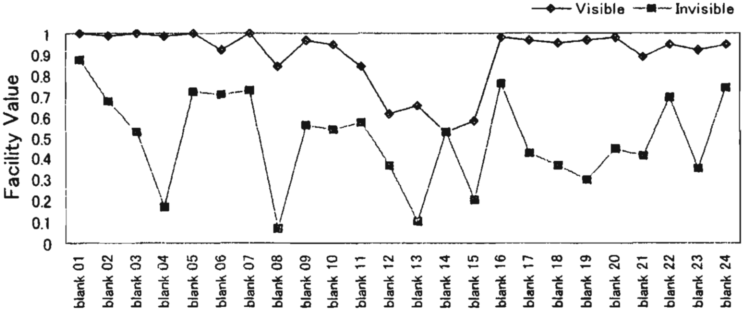

Figure 1. Comparison of the visible and the invisible facility values of the same gap

3.2. Item Discrimination

Examined next was item discrimination. Employed as a discrimination index of each item was the point-biserial correlation between an item score and the composite score, i.e., the sum of the numbers correct on the visible test (k = 24), the invisible test (k = 24), the achievement c-test (k = 25), and the proficiency c-test (k = 25). Since the composite score (k=98) was our best estimate of the candidate's ability, the coefficient should indicate the extent to which each visible/invisible item succeeded in differentiating between stronger and weaker candidates.

The descriptive statistics of item discrimination are summarized in Table 3. The mean was .296 for the visible items and .457 for the invisible items. The minimum for the visible items was naturally .000, due to the fact that four of them had a facility value of 1.000 as mentioned above. The range (.406) and the SD (.150) were both smaller for the invisible items than for the visible items (.650 and .207, respectively). A t-test indicated that the difference in the means was highly significant (t = 3.22, p = .004, two-tailed).

Table 3. Descriptive statistics of item discrimination values

	k	N	Mean*	SD*	Max	Min
Visible	24	88	.296	.207	.650	.000
Invisible	24	88	.457	.150	.647	.241

Note: The means and the SDs were computed based on Fisher-z-transformed values.

The distributions are graphically represented as Figure 2. It is clearly seen that item discrimination values for the invisible items clustered together much more closely in the right-hand side of the possible distribution range than their visible counterparts. According to Ikeda's (1992) criterion, items with discrimination values of .4 or higher are considered "good", those with values of .3 or higher and lower than .4 are considered "acceptable", those with values of .2 or higher and lower than .3 "have room for improvement", and those with values lower than .2 are "poor items that need to be discarded or completely rewritten". Based on this classification, of our visible items, only 33% were good, 17% were acceptable, 17% needed improvement, and as many as 33% needed to be discarded. On the other hand, our invisible items performed much more favorably, with as many as 63% in the "good" category, 25% in the "acceptable" band, 12.5% needed improvement, and none needed to be discarded.

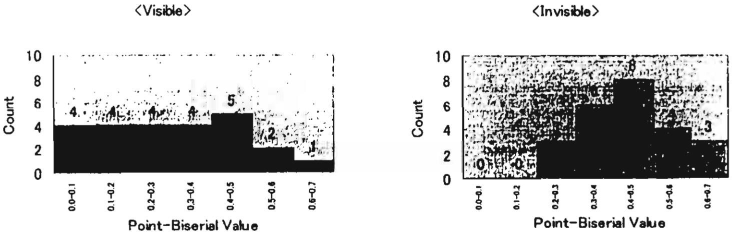

Figure 2. Distribution of discrimination: Visible and invisible items

Of our next interest was whether visible items with higher (or lower) discrimination made higher (or lower) invisible items. The discriminations for the 24 item pairs based on the same gaps are shown in Figure 3. Apparently, it is not the case that well-discriminating visible items make well-discriminating invisible items or non-discriminating visible items correspond to non-discriminating invisible items. Lack of such correspondence is typically shown by the fact that the four gaps that produced a discrimination value of 0.00 when they were visible (gaps 01, 03, 05, and 07) were transformed to highly discriminating items when they were invisible (point-biserials were .48, .58, .41, and .64, respectively) and, conversely, the two most highly discriminating visible gaps (gaps 08 and 15) were changed into the two least discriminating invisible gaps. In fact, the Pearson correlation was not significant ($r = -0.239$, n.s).

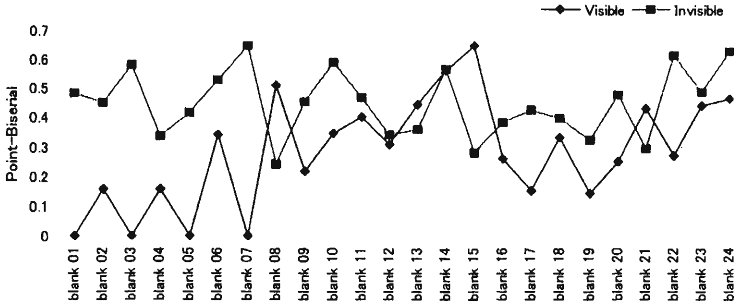

Figure 3. Comparison of the visible and the invisible discriminations of the same gap

Investigated next was the relationship between item facility and item discrimination. It is known that, other things being equal, items with very high or very low facility values tend to make non-discriminating items and those with mediocre facility will make highly discriminating items, resulting in an inverted U-shaped curve when plotted. Figure 4, a facility-discrimination scatter plot, confirms this curvilinear relationship.

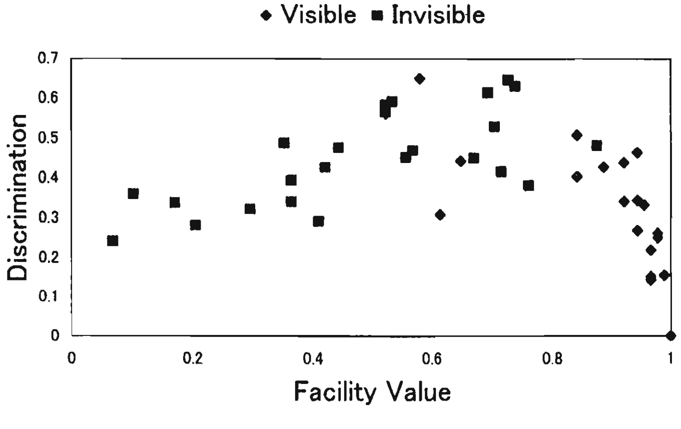

Figure 4. Item facility-discrimination scatter plot of visible and invisible items

It can be seen that the visible items are concentrated around the bottom-right area of the curve gradually spreading over to the top-middle area, whereas the invisible items are distributed over a larger area ranging from the middle-left to high-top positions. One question worth asking is whether the obtained higher discriminations of invisible items are mere artifacts of their facility values. Were the visible items generally non-discriminating simply because they were generally too easy, and were the invisible items, on the contrary, better discriminating just because they were generally of more appropriate facility values? Or was there something other than appropriate facility values at play?

To answer these questions, a multiple regression analysis was attempted. The predicted variable was item discrimination value. The predictor variables were "item-facility distance" (IF-DIST) and "gap visibility" (VISIBILITY). It is known that other things being equal, the facility value that produces the highest discrimination is the mid-point between the mean success probability of random guessing and 1.00. For example, for a four-option multiple-choice item, that facility value is the mid-point between .25 and 1.00, which is .625. For our visible items, which involved five-against-five matching, the optimum value was the mid-point between .20 and 1.00, or .60. On the other hand, the mean success probability of random guessing for our invisible items could virtually be considered zero since the chance of random guessing the gap correctly was only one out of all the number of words in that portion of text. Hence, the optimum facility value for invisible items was the mid-point between 0.00 and 1.00, or .50. Therefore, IF-DIST of an item was defined as its facility value's distance from .60 in the case of a visible item, and as that from .50 in the case of an invisible item. VISIBILITY was a categorical variable indicating whether the gap of the item was visible or invisible. To include this as a regression term, VISIBILITY was coded as 1 for a visible item and 0 for an invisible item. This was a what is known as a dummy variable.

The model fitted to the data was as follows:

$$\text{Item discrimination} = \beta_0 + \beta_1 \text{IF-DIST} + \beta_2 \text{VISIBILITY}$$

where

IF-APP: tem-facility distance, as defined above

VISIBILITY: 1 for a visible item; 0 for an invisible item

This model resulted in an R of .698, an R-squared of .487, and an adjusted R-squared of .464. The model overall was highly significant as shown by the F-value (=21.35) and the p-value (= .000) in Table 4.

Table 4. Analysis of variance

	Sum of Squares	df	Mean Square	F	Sig
Regression	0.695247	2	0.347624	21.35126	0.000
Residual	0.732653	45	0.016281		
Total	1.4279	47			

The focus of our interest was the significance of the partial regression coefficient of VISIBILITY. As Table 5 shows, the p-value was .055, not reaching the commonly employed significance of .05.

Table 5. Regression coefficients

	Unstandardized		Standardized		
	B	Std. Error	Beta	t	Sig.
(Constant)	0.582	0.038		15.271	0.000
VISIBILITY	-0.080	0.041	-0.233	-1.966	0.055
IF-DIST	-0.725	0.152	-0.565	-4.774	0.000

Hence, the regression model obtained, using unstandardized coefficients, was:

$$\text{Item Discrimination} = 0.582 + (-0.725) * \text{IF-DIST} + (-0.080) * \text{VISIBILITY}$$

Both the regression coefficients were in the negative. This means that (1) visibility condition being held constant, item discrimination gets lower when item facility distance gets larger; and that (2) item facility distance being held constant. item discrimination is lower when visibility condition is 0 (i.e., the item is invisible) than when it is 1 (i.e., the item is visible).

However, since the significance of the coefficient for VISIBILITY was only 0.055, we do not have much confidence in the stability of this model.

3.3. Test Reliability

Next, we turned to comparing test reliability. The Cronbach's alpha for the visible-gap test was .7912, while that for the invisible-gap test was .8658. The higher reliability of the invisible test was predictable after we learned of generally higher discrimination of invisible items and the larger standard deviation in facility values of invisible items. Item discrimination and score standard deviation are the two most important factors affecting test reliability. According to Spearman-Brown prophecy formula, the visible-gap test needs to be lengthened 1.70 times in order to achieve the reliability attained by the invisible-gap test.

3.4. Test Validity

Lastly we turned to validity analyses. To recapitulate, our interest was in the extent to which the scores on the visible and the invisible tests, both based on the same material covered in class, would predict the scores of c-test A (based on a familiar passage) and c-test B (using an unfamiliar passage). Before presenting the criterion-relatedness results, descriptive statistics and reliabilities of c-tests are shown in Table 6. The mean of c-test A was much higher than that of c-test B, probably because the former was based on a "covered" passage. Alphas were .7736 for c-test A and .6666 for c-test B. The lower value for the latter may be attributed to its smaller standard deviation. Combined, the c-test total's (k = 50) alpha exceeded .8. The c-test A score was our best estimate of the general degree to which the students processed the form and the meaning of reading materials covered in class, and the c-test B score was our best estimate of the degree to which the students were able to process new materials of comparable readability. The c-test composite score was our best estimate of the candidate's general reading ability. It should be noted that the term "best" in this context does not imply that the estimates are accurate to a great degree in absolute terms but simply that they are the best among the available data.

Table 6. Means, standard deviations, and reliabilities of c-tests

	C-test A	C-test B	C-test Total
k	25	25	50
N	88	88	88
Mean	17.84	10.89	28.72
SD	3.91	3.55	6.71
alpha	.7736	.6666	.8283

Validity results are summarized in Table 7. Figures in the columns with the heading "r" are Pearson correlations and those in the columns headed with "r^2" are r-squared, or coefficients of determination. It can be seen that c-test A correlated more strongly with the invisible test (r = .694, p < .01) than with its visible counterpart (r = .534, p < .01). As the r-squared values indicate, the invisible score predicted 48.2% of the c-test A variance while the visible score accounted only for 28.5%. Likewise, c-test B had a stronger relationship with invisible-gap filling (r = .510, p < .01) than with visible-gap filling (r = .399, p < .159); the percentages of c-test variance accounted for were 26.1 and 15.9, respectively. When the criterion was the c-test composite score, the results were quite similar; the r with the invisible score was .674 (p < .01) and that with the visible score was .523 (p < .01).

Note that the point here was not whether the invisible test's correlation was strong enough to make it a practical predictor of the criterion, which was not our concern. What was of interest was whether the strength of relationship between c-test and gap filling significantly changed depending on gap visibility conditions. Therefore, the differences between non-independent correlation pairs were tested for significance, the results of which are shown in the right-most column of the table. Though the difference for c-test B failed to reach significance (p = .10), those for c-test A and c-test total both did (p < .01). Therefore, it was indicated that the invisible-gap filling test was a significantly better predictor of the achievement c-test score and the c-test total score. In the case of the proficiency c-test score, the invisible score was a better predictor than the visible score, though the difference was not large enough to reach significance.

Table 7. Validity coefficients, coefficients of determination, and t-test results

	Invisible		Visible		t-test results
	r	r^2	r	r^2	
C-test A	.694**	.482	.534**	.285	t = 2.85, p < .01 (2-tailed)
C-test B	.510**	.261	.399**	.159	t = 1.65, p = .10 (2-tailed)
C-test total	.674**	.455	.523**	.273	t = 2.64, p < .01 (2-tailed)

It should be noted that these observed coefficients are all artifacts of less-than-perfect reliability of each measure. By correcting for attenuation, one can estimate the strength of relationship that should be observed when the measuring instruments were perfectly reliable. Table 8 shows validity coefficients, coefficients of determination when the two gap-filling tests and two c-tests are all corrected for attenuation, and t-test results based on those corrected values. Naturally, the coefficients in Table 8 are all higher than their corresponding coefficients in Table 7, due to the improved reliabilities. It can be seen that the pattern of values for the invisible scores being higher than those for the visible scores are consistently carried over even after correction for attenuation is applied. An important observation is that the invisible-visible difference in predicting power when c-test B is the criterion is now significant (p < .01). Hence it is indicated that lack of significance in the actual coefficients difference was probably due to the low reliability of c-test B.

Table 8. Validity coefficients and coefficients of determination when corrected for attenuation, and t-test results

	Invisible		Visible		t-test results
	r	r^2	r	r^2	
C-test A	.848**	.719	.683**	.466	t = 6.89, p < .01 (2-tailed)
C-test B	.672**	.451	.550**	.302	t = 3.45, p < .01 (2-tailed)
C-test total	.796**	.634	.645**	.417	t = 5.35, p < .01 (2-tailed)

4. DISCUSSION

This study examined comparable visible- and invisible-gap filling tests based on the same familiar passage as two possible forms of reading-centered achievement tests. The focus of the analysis was to see how removing gap indicators would affect the psychometric properties of the test.

The first observation was that the removal made the test more difficult. This may appear to be stating the obvious, but there is more to it. First, making the gaps invisible made the test more appropriate for the target group in terms of difficulty. A multiple-matching visible-gap filling test, such as the visible test in this study, tends to be too easy when the passage is a familiar one. The mean percentage correct for the visible test was 89.1, and four items were correctly responded to by all the test-takers.

Assuming that the test was designed to be a norm-referenced one that differentiates among students, the visible version was clearly a failure. The invisible counterpart, on the other hand, had a mean percentage correct of 49.1, a much more appropriate value.

It was not that removing the gap indicators simply made all the items more difficult by similar degrees. To be more precise, the removal caused all but one item to become more difficult by widely varying degrees. The difficulties of the visible items ranged from .523 to 1.000, but when the gaps were invisible, the items varied from quite easy (IF = .875) to very difficult (IF = .068). It is a virtue for a test to have items at a wide range of difficulty levels in that such a test can differentiate among candidates both at higher and lower areas on the ability continuum. In this regard, it can be maintained that making the gaps invisible contributed to making the test more appropriate for test-takers distributed over a wider proficiency range.

The second finding was that when the gaps were invisible, item discriminations were markedly higher. When the gaps were visible, as many as 50% of the items were poorly discriminating, but when the gaps were invisible, the percentage of the items that needed editing was only 12.5%. This could be considered a rather impressive betterment. This should be attributed to the introduction of an additional task of locating the gaps themselves.

One interesting observation with regard to item discrimination was that discriminations of visible and invisible items were not correlated. This could be interpreted as indicating that a crucial factor in discriminatory power of an invisible item is that it requires careful reading to locate the gap itself. That is, the discriminatory power of an invisible item probably resides primarily in the gap-locating, rather than in the gap-filling, phase.

By the regression analysis, we attempted to factorize discriminatory power of invisible items into two components: that which derives from appropriate facility values and that which comes from the gap invisibility itself. The obtained model did not provide a clear-cut answer, though. The model indicated that discriminations of the invisible items are clearly attributed to their item facility appropriateness. There was also some indication that even after item facility appropriateness factor is accounted for, the gap invisibleness by itself is making a positive contribution to discrimination. However, since p-value of that tendency was only .10, this is little more than a speculation for the time being. One way of checking the stability of a regression equation is to divide the sample into two subsamples, derive an equation using the same predictors from each subsample, and conduct an F-test to check the comparability of the two models (Crown 1998: 47-48). This could not be attempted in this study due to time constraint, but is an analysis that needs to be conducted in a follow-up study.

Given more appropriate facility values and higher discrimination, the higher reliability observed for the invisible version was a natural consequence. Even so, it is an impressive fact that simply removing noticeable gap indicators from a multiple-matching gap-filling test made it markedly more reliable, to the extent that the original test needs to be lengthened 1.7 times to be equally reliable

The validity analysis revealed straightforward results as well. The invisible version was simply more valid as an achievement measure as well as a proficiency measure. Since this study compared invisible- and visible-gap filling, the latter of which can be termed a multiple-matching rational cloze procedure, one might be tempted to regard invisible-gap filling as just another variant of a cloze test. We do not share this view. Whether mechanical or rational, the essence of the cloze procedure is that it requires the reader to provide possible words to fit explicitly presented gaps. On the other hand, as is indicated by the lack of correlation in item discriminations between the visible and invisible version of the tests in this study, the essence of invisible-gap filling seems to reside in its gap invisibility, rather than in the fact that the gap needs to be filled. In order to locate a gap, the reader needs to fully understand the parts surrounding it, in terms of both form and meaning. Only when the reader is reading with full comprehension can they notice a gap that needs to be filled by some word. It should follow that whether one can correctly locate a gap reflects whether one is reading with full comprehension the parts immediately preceding and following it. This seems exactly what makes an invisible gap filling test a more valid reading measure than a visible one.

A possible doubt about this item type may be: "When we read, we do not look at every word. The passage is perfectly comprehensible even with these "gaps". Even a native speaker may not notice some of them. Why bother paying attention to such gaps when you can understand everything? After all, comprehending the content is what reading is all about." In our view, this comment reveals a misconception about the nature of reading and the objective of testing reading. First, the popular notion that we "do not look at every word" when we read is almost a myth. As pointed out by Stanovich (1991), research consistently shows that the vast majority of content words as well as the majority of function words in text receives a direct visual fixation (Ehrlich & Rayner 1981; Just & Carpenter 1980; 1987; Perfetti 1985). The sampling of visual information in reading is much more dense than is commonly believed. Second, the purpose of a reading test is to measure how well the candidate can read, not necessarily to have reader read as they would in a non-test-taking situation. Even if a native speaker *will not* notice some invisible gaps when they are reading 'naturally', (that is, somewhat carelessly), they certainly *can* locate all of them when they are required to pay attention. The situation is quite different with a non-native reader, who sometimes *cannot*, as opposed to *will not*, locate the gaps. The objective of the test is to find out who can, and who cannot, locate them to what extent, thereby enabling us to estimate their underlying reading ability. The fact that some readers *will not* locate them when reading 'naturally' is essentially irrelevant to the legitimacy of the test.

Before concluding, one limitation of the invisible-gap filling used in the present study needs to be pointed out. In retrospect, there may have been a drawback in providing the word groups to fill in the invisible gaps. The originally assumed test-taking behavior is: the candidate (1) reads the text with comprehension until (2) they notice a grammatical or semantic gap, (3) looks down at the group of possible filler words, (4) chooses one, and (5) goes on reading the text again. However, a more test-wise candidate may first (1) look at the group of filler words, (2) make some predictions concerning where these words should belong, (3) and then browse through the text testing the correctness of those prediction. This item tackling tactics may have nothing wrong in itself, but if it is adopted by only a part of the candidates, it may bring unpredictable variables into play, complicate the test-taking behavior of the group as a whole, and thereby lowers of the test reliability.

To avoid this complication, there seems to be only one way out, which is to remove the filler words from the test. After removing them, there are two courses of action to take: one is to require candidates to provide, as opposed to choose from among options, possible filler words themselves, and the other is not to require them to do so at all. The latter option results in a test in which the candidate only indicates locations of perceived gaps in the text, without answering what words are likely to fill those gaps. The first type, a constructed-response invisible-gap filling test, may be even more sensitive than the multiple-matching counterpart to proficiency differences, since it can now differentiate between those who can choose but cannot provide filler words and those who can also provide them without options.

The second type, an invisible-gap *locating* test, is attractive as well. One may wonder if a test that requires candidates only to indicate locations of the gaps without answering what words should fill those gaps will be equally reliable and valid as the one for which fillers must be chosen, but the results of this study indicates that it will be. Indirect evidence is that discriminations of visible and invisible items did not correlate with each other. This could, as maintained above, be interpreted as indicating that discriminatory power of invisible gap filling is mostly in the gap locating, not in the gap filling, phase. More direct evidence is that a correctly located gap nearly always meant a correctly filled one. Since locating a gap seemed always more difficult than filling the gap once it was located, someone able enough to spot an invisible gap almost always reached the final correct answer. In this regard, gap-filling phase was virtually redundant. It follows that a test that only requires the candidate to indicate the location of a gap may have served almost the same purpose.

Note and Acknowledgement

This is an elaborated version of the paper read at the JLTA Conference 2002, held at Tokyo Keizai University on October 2, 2002. The author thanks Professor Jeffrey Hubbell at Hosei University for the insightful comments he made on the orally presented version.

Works Cited

Crown, W. H. *Statistical Models for the Social and Behavioral Sciences: Multiple Regression and Limited-Dependent Variable Models.* Westport, CT: Praeger Publishers, 1998.

Ehrlich, S. F. & Rayner, K. "Contextual Effects on Word Perception and Eye Movements During Reading." *Journal of Verbal Learning and Verbal Behavior* 20 (1981): 641-655.

Hughes, A. *Testing for Language Teachers*, 2nd ed. Cambridge: Cambridge University Press, 2003.

Ikeda, H. *Tesuto no Kagaku [Scientific Testins].* Tokyo: Nihon-Bunka-Kagakusha, 1992.

Just, M. A., & Carpenter, P. A "A Theory of Reading: From Eye Fixations to Comprehension." *Psychological Review* 87 (1980): 329-354.

Just, M. A., & Carpenter, P. A. *The Psychology of Reading and Language Comprehension.* Boston: Allyn & Bacon, 1987.

Meyer M. & Felton, R. "Repeated Reading to Enhance Fluency: Old Approaches and New Directions. *Annals of Dyslexia.* 49 (1999): 283-306.

Perfetti, C. A. *Reading Ability.* New York: Oxford University Press, 1985.

Samuels, S. Jay. "Reading Fluency: Its Development and Assessment." *What Research Has to Say About Reading Instruction.* Eds. A. Farstrup & S. J. Samuels. Newark, DE: International Reading Association, 2002. 166-183.

Shizuka, T. *Eigo Testo Sakusei No Tatujin Manyuaru* [Writing English language tests: A manual for teachers]. Tokyo: Taishukan, 2002a.

Shizuka, T. *Reading Communicator: Read and Think About 20 Current Topics.* Tokyo: Sanshusha, 2002b.

Stanovich, K. E. "Changing Models of Reading and Reading Acquisition." *Learning to Read: Basic Research and Its Implications.* Eds. L. Rieben & C. A. Perfetti. Hillsdale, NJ: Lawrence Erlbaum Associations, 1991. 19-31.

Taguchi, E. "The Effects of Repeated Readings on the Development of Lower Identification Skills of FL Readers." *Reading in a Foreign Language* 11 (1997): 97-119.

Appendix: The tests used in the study

<The Invisible-Gap Filling Test>

[1] 次の一続きの５つの文章の中からはそれぞれ５つの語が抜けています。まず、（１）語が抜けている５つの箇所の直前の語を書き、（２）次に、それぞれの場所に入るべき語を選んで書きなさい。<1×25>

	直前の語	入る語
If there is a picture of you somewhere on the Web, someone with common image-editing software, and a little time and expertise, can easily paste your onto someone else's body and repost the image. "Faked" nude photos are more commonly associated with supermodels and celebrities, everyday individuals can also fall victim, as Mark Hall, a warehouse manager in Connecticut, and his wife Nancy learned. Hall, 33, and his wife put a picture of her on Rankpeople.com, a Web site that allows to look at pictures of people and rate their appearance a one-to-10 scale. Within a day or two, Nancy called her husband at work to tell him someone had altered the image so that appeared naked, reposted it on the site, and anonymously e-mailed her Rankpeople.com account.		

(抜けている語　/　on　/　head　/　she　/　users　/　but)

"It looked real," Hall said. "You couldn't tell it's a fake," agreed Lt. James Cetran, investigated the case once the Halls complained to police. The Halls e-mailed the creator of the image, asking him or her to it, but instead they received an e-mail said, "I might remove it if you send me more pictures of yourself." After several days, the Halls were able to get Rankpeople.com to remove the image, but the experience left them. "What somebody at her workplace goes and sees it?" Hall asks. "The public side — that's what bothered us."		

(抜けている語　if　/　that　/　who　/　uneasy　/　remove)

The easiest recourse for victims may be to the Internet service provider hosting the photo to remove it. Rankpeople.com notes that about photos are reviewed by staff members, and that the pictures are automatically removed if they repeated criticism. "Rankpeople takes every to make sure that nothing up there is," says Tabitha Sturm, a company spokeswoman.		

(抜けている語　indecent　/　ask　/　generate　/　complaints　/　precaution)

The Halls' case far from unique, says Mark Rasch, a cyberlaw for Predictive Systems and former Justice Department computer crimes prosecutor. "This happens a lot," he says, particularly in cases of coworker disputes and divorce cases. "I mad at you, I want to get back at you, so what can I do?" he says hypothetically. "This is one option." Another Internet expert cautions, however, against the risk of someone altering and reposting of a private individual.		

(抜けている語　get　/　expert　/ exaggerating　/　is　/　pictures)

"It does happen," says Steven Jones, a University of Illinois, Chicago communications professor also serves as president of the Association of Internet Researchers. But, he says, "It's they should probably worry about as much as they worry about the street." Furthermore, he says, people must realize that they put something on the Web, it is next to impossible to people from altering it and reposting it. "It's the nature of the beast," he admits.		

（抜けている語　who / crossing / prevent / something / once）

<The Visible-Gap Filling Test>

[２] 次の５つの文章の中からはそれぞれ５つの語が摘出されています。それぞれの場所に入るべき語を選んで書きなさい。<1×25>

If there is a picture of you somewhere on the Web, someone with common image-editing software, and a little time and expertise, can easily paste your ( ) onto someone else's body and repost the image.

"Faked" nude photos are more commonly associated with supermodels and celebrities, ( ) everyday individuals can also fall victim, as Mark Hall, a warehouse manager in Connecticut, and his wife Nancy learned.

Hall, 33, and his wife put a picture of her on Rankpeople.com, a Web site that allows ( ) to look at pictures of people and rate their appearance ( ) a one-to-10 scale.

Within a day or two, Nancy called her husband at work to tell him someone had altered the image so that ( ) appeared naked, reposted it on the site, and anonymously e-mailed her Rankpeople.com account.

（抜けている語　on / head / she / users / but）

"It looked real," Hall said. "You couldn't tell it's a fake," agreed Lt. James Cetran, ( ) investigated the case once the Halls complained to police.

The Halls e-mailed the creator of the image, asking him or her to ( ) it, but instead they received an e-mail ( ) said, "I might remove it if you send me more pictures of yourself."

After several days, the Halls were able to get Rankpeople.com to remove the image, but the experience left them ( ). "What ( ) somebody at her workplace goes and sees it?" Hall asks. "The public side — that's what bothered us."

（抜けている語　if / that / who / uneasy / remove）

The easiest recourse for victims may be to ( ) the Internet service provider hosting the photo to remove it. Rankpeople.com notes that ( ) about photos are reviewed by staff members, and that the pictures are automatically removed if they ( ) repeated criticism. "Rankpeople takes every ( ) to make sure that nothing up there is ( )," says Tabitha Sturm, a company spokeswoman.

（抜けている語　indecent / ask / generate / complaints / precaution）

The Halls' case ( ) far from unique, says Mark Rasch, a cyberlaw ( ) for Predictive Systems and former Justice Department computer crimes prosecutor. "This happens a lot," he says, particularly ( ) cases of coworker disputes and divorce cases. "I (   ) mad at you, I want to get back at you, so what can I do?" he says hypothetically. "This is one option."

Another Internet expert cautions, however, against the risk of someone altering and reposting (   ) of a private individual.

（抜けている語　get　/　expert　/　is　/　pictures　/ exaggerating)

"It does happen," says Steven Jones, a University of Illinois, Chicago communications professor (   ) also serves as president of the Association of Internet Researchers. But, he says, "It's (   ) they should probably worry about as much as they worry about (   ) the street." Furthermore, he says, people must realize that (   ) they put something on the Web, it is next to impossible to (   ) people from altering it and reposting it. "It's the nature of the beast," he admits.

（抜けている語　who　/　crossing　/　prevent　/　something　/　once)

<The C-Tests>

[3]　次の２つの文章（お互いに独立しています）の単語は１語おきに後ろ半分がアンダースコア"_"で置換してあります。"_" ひとつが１文字に相当します。文字数をヒントに語尾などにも注意して正確に復元しなさい。<1×50 >

<A>

Medical applications of gene therapy are at a primitive stage. Little i_ understood ab___ the implic______ of intro______ foreign ge___ into _ human bo__, and s_ any u__ for impr_____ athletic perfo______ now wo___ be consi_____ dangerous a__ unethical. B__ the hu___ genome h__ been map___ out a__ the techn_____, however imma____, is evol____ rapidly. Athl____, who a__ often ea___ for an edge in competition, are not likely to wait for medical science to perfect gene therapy.

<B>

Much of Asia sees Japan as a country with a split personality, a hard-to-understand culture that inspires contradictory sentiments. The wo___ of t__ stereotypes, warm_____ Japan -- t__ country th__ invaded i__ neighbors, for___ men t_ toil i_ munitions fact_____ and wo___ to wo__ as sex sla___ for i__ soldiers -- i_ still al___ in n_ small pa__. That i_ primarily bec____ Japan h__ done su__ a po__ job exorc_____ the o__ demons. The apologies came too late and are too feeble. Japan didn't acknowledge it forced Korean, Taiwanese and Filipino women into sex slavery until 1993.

Note: The test is reproduced here as it was used in the study, including the faults in the Invisible test. Note that *case*, the correct word to fill in the third gap in the fourth block ". . . in (   ) of . . .", was not included in the options. (Instead, *exaggerating*, which did not fit any of the gaps in that part, was. In fact, *exaggerating* was missing from the text, after *against* and before *the risk*.) Due to this complication, the item was dropped from the analysis, together with its visible counterpart.