OPEN ACCESS

Vocabulary Learning and Instruction
ISSN 2981-9954
Volume 13, Number 2 (2024)
https://doi.org/10.29140/vli.v13n2.2126

Castledown

Estimating the Impact of Local Item Dependency in a Test of Second Language Reading Comprehension

Tim Stoeckel
University of Niigata Prefecture, Japan

Liang Ye Tan
St. Andrew's University, Japan

Hung Tan Ha
Victoria University of Wellington, New Zealand

Nam Thi Phuong Ho
University of Economics Ho Chi Minh City (UEH), Vietnam

Tomoko Ishii
Meiji Gakuin University, Japan

Young Ae Kim
Kyoto Seika University, Japan

Chunmei Huang
Kyorin University, Japan

Stuart McLean
Kindai University, Japan

Abstract

Local item dependency (LID) occurs when test-takers' responses to one test item are affected by their responses to another. It can be problematic if it causes inflated reliability estimates or distorted person and item measures. The cued-recall reading comprehension test in Hu and Nation's (2000) well-known and influential coverage–comprehension study may have been susceptible to LID because of its very high sampling rate of idea units from the test's reading passage. To explore this possibility, we examined Yen's Q_3 statistics in a dataset comprising responses from 63 learners who took Hu and Nation's test. Using Q_3 values of ≥ .30 to flag possible instances of LID, 14 item pairs were identified as being locally dependent. To remove LID, items with high Q_3 values were combined into polytomous super-items. Using Rasch measurement, a comparison of the data with and without the effects of LID indicated that LID only modestly inflated reliability estimates and that there was a negligible effect on the ordering of persons by ability. Thus, while LID was present, it had only a small impact on how test scores might be interpreted in a study like Hu and Nation's.

Keywords: local item dependency, Rasch measurement, partial credit Rasch model, Hu and Nation.

Local item dependency (LID) occurs when test-takers' responses to one test item are affected by their responses to another (Fan & Bond, 2019). LID has various causes, but one situation that is particularly conducive to LID is when multiple test questions share a single stimulus such as a reading passage. Although minor manifestations of LID may be unproblematic (Aryadoust et al., 2021), more severe cases can artificially inflate reliability estimates and distort person and item measures. Thus, test data should be examined for LID, and, when present, its effects on measurement should be determined. This study examined the effects of LID in a reading comprehension test from an influential coverage–comprehension study.

The Hu and Nation Reading Test

Coverage–comprehension research (McLean, 2021) aims to define the relationship between the percentage of tokens in a text that are known (i.e., coverage) and the degree to which the text is understood (i.e., comprehension). In one such study, Hu and Nation (2000) identified 98% as an important coverage level for independent reading. This study has been immensely influential in L2 English pedagogy: Its 98% coverage figure has been used to estimate the vocabulary sizes needed for independent reading in different genres (Nation, 2006) and is routinely utilized as a threshold of importance in research (e.g., Laufer & Cobb, 2020).

Despite the impact of Hu and Nation's study, its cued-recall reading test (Appendix A) may have been susceptible to LID. To assess the understanding of numerous idea units in their 620-word passage, those researchers developed an instrument comprising 27 prompts that elicited 44 separate pieces of information. This high sampling rate may have unavoidably caused the responses of some items to be dependent on others. For example, it would be nearly impossible to correctly answer Q12 unless the answer to Q11 were also known:

> 11. What were her feelings when she went through the house?
>
> 12. Why did she feel this way?

Nonetheless, the thoroughness of Hu and Nation's approach was appropriate for coverage–comprehension research. Because the aim of such studies is to determine the relationship between knowledge of the lexis and understanding of the ideas in a text, when either of those constructs is undersampled, the error in estimating that relationship increases. Put another way, if comprehension is measured by assessing the understanding of just a few ideas in a text, only the lexis needed to understand those particular ideas is required to succeed.

Detecting LID and Quantifying its Effects

There are various methods for detecting LID, with perhaps the most common approach in L2 research being the inspection of Yen's (1984) Q_3 statistics. Q_3 is simply the correlation between the raw residuals for each pair of test items. It quantifies the amount of common variance in the response strings for two items after the shared variance accounted for by test-takers' ability is conditioned out. Simulation research has found that there is no single critical Q_3 value for identifying LID (Christensen et al., 2017), but Aryadoust and colleagues (2021) have suggested that item pairs with Q_3 values of $< .30$ are unlikely to distort measurement.

When potentially impactful residual correlations are present, their effect on measurement can be estimated by comparing analyses of (a) the original test data with (b) the data with the residual correlations removed. In Rasch measurement, residual correlations can be removed by combining correlating items into polytomous super-items and re-analyzing the data with a polytomous Rasch model (Baghaei, 2008; Marais & Andrich, 2008).

Purpose

The purpose of this study was to see whether LID occurs in the Hu and Nation cued-recall reading test and, if so, to quantify its impact on estimates of reliability and person measures.

Method

The participants were 63 L2 learners of English. They included university students in Vietnam (n = 31), Japan (n = 23), and China (n = 9). Although standardized test data were unavailable, we estimate that their levels in the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) to have been mostly A2 and B1, with perhaps a few learners at the A1 and B2 levels. The instrument was Hu and Nation's cued-recall reading test, which consists of 27 questions that elicit 44 distinct idea units (Appendix A). For clarity, hereinafter, the 27 questions are referred to as *prompts* and the 44 elicited concepts as *items*. Following Hu and Nation, the participants read the text and then answered the questions without access to the text. The test prompts and elicited responses were in the participants' L1. Responses to the 44 items were marked dichotomously[1] (see Appendix B for the key). There were six markers, two for each of the participants' L1 backgrounds. All markers were applied linguists, were highly proficient in English, and shared an L1 with the participants whose responses they marked. Based on guidelines by Landis and Koch (1977),

[1]Hu and Nation used weighted scoring. We conducted two sets of analyses, one with and one without the item weights used by Hu and Nation. As the two approaches yielded virtually identical findings, the more parsimonious approach with no weighting is reported here.

Cohen's kappa estimates of inter-rater reliability were "substantial" in all cases: Vietnamese κ = .745 [.709, .782]; Japanese κ = .757 [.732, .782]; Chinese κ = .670 [.596, .745].

The data were analyzed three times. In Analysis 1, the 44-item test was analyzed with the dichotomous Rasch model. In Analysis 2, items with a shared stimulus (i.e., prompt) were combined to form polytomous super-items, which created a 27-item test, and the data were analyzed with Masters' (1982) partial credit Rasch model. To form a super-item, scores on two or more of the initial 44 items were simply summed (Marais & Andrich, 2008). For example, whereas the three pieces of information elicited by the first prompt each received a score of 0 or 1 in Analysis 1, they were combined into a single, polytomous super-item in Analysis 2, with a possible score ranging from 0 to 3. As some high residual correlations remained in the second analysis, in Analysis 3, the items were further consolidated. This time, eight polytomous super-items were formed, each representing one contiguous section of the reading passage (see Appendix B), and Masters' partial credit model was again employed. Winsteps version 5.2.2 (Linacre, 2022) was used for all analyses. Q_3 values of ≥ .30 were considered to represent high residual correlations.

Analysis 1 represents test data in the form that was used by Hu and Nation, whereas Analysis 3 represents the same data with the effects of LID removed. Thus, differences between these two analyses denote the effects of LID. Hu and Nation's test has been used in research to differentiate learners for purposes of regression analysis; hence, Rasch person reliability coefficients were compared, and raw scores were correlated with the Rasch person measures from Analyses 1 and 3. Moreover, because LID has been associated with changes in the distribution of person estimates (Marais & Andrich, 2008), scatterplots of person measures with and without the effects of LID were examined.

In support of open science, the three complete datasets and Winsteps control files can be found at https://osf.io/375wq/?view_only=41e3f67a3a0042a382ae7e592dffd31f.

Results

Q_3 Values

Table 1 displays the item pairs with Q_3 values ≥ .30. In Analysis 1 in particular, many of the item pairs with high residual correlations were either adjacent items or items belonging to the same prompt. Such cases are the most likely to indicate true item dependency because the response to one item may indeed depend on that of another. This was clearly the case, for example, for the adjacent items 17 and 18, which had the highest residual correlation (Q_3 = .75):

17. What did she tell them to do?

18. Why did she tell them to do this?

For Analysis 2, recalling that only items sharing a test prompt were combined, the highest Q_3 value of .75 remained because it involved separate prompts (17 and 18).

Table 1 *Standardized Residual Correlations of ≥ .3*

Analysis 1			Analysis 2			Analysis 3		
Item	Item	Q_3	Item	Item	Q_3	Item	Item	Q_3
17	18	.75	17	18	.75		(none)	
15_1	15_2	.70	10	14	.32			
13_4	13_5	.67	7	8	.31			
13_1	13_3	.53	17	19	.31			
7_2	8	.42	14	17	.30			
6_3	7_1	.40	14	18	.30			
13_1	13_2	.39						
9_1	12_1	.37						
4	13_4	.33						
2_1	16	.31						
17	19	.31						
14	17	.30						
14	18	.30						
9_1	14_1	.30						

Note: For Analysis 1, the item numbers correspond with the number of the test prompt, and prompts that required more than one piece of information have underscores followed by values to show this. For Analysis 2, this was unnecessary because prompts requiring multiple responses were combined into super-items.

Table 2 *Range, Reliability, and Separation Indices for Person Measures*

Analysis	Range	Reliability		G	Strata
		Real	Model		
1: 44-item, dichotomous	5.25	.87	.88	2.72	3.29
2: 27-item, partial credit	5.14	.86	.88	2.67	3.23
3: 8-item, partial credit	4.68	.82	.85	2.39	2.85

Note: G = person separation (model). Strata = (4G – 1)/3.

Conversely, the high inter-prompt correlations for Prompts 13 and 15 in Analysis 1 were removed in the formation of super-items for Analysis 2. Table 1 indicates that the item pairs with Q_3 values greater than .30 in Analysis 2 were either for adjacent items (e.g., 17 and 18) or items from approximately the same part of the passage. These statistics guided how items were merged for Analysis 3. No items had Q_3 values of ≥.30 in Analysis 3.

Person Reliability and Separation

Table 2 shows the range, reliability, and separation indices for the Rasch person measures for Analyses 1–3. When strong residual correlations were removed by creating polytomous super-items in Analyses 2 and 3, the range of person measures was reduced. Marais and Andrich (2008) also observed this phenomenon. As this made it harder to distinguish test-takers according to the tested construct, the person reliability estimates and separation indices worsened. These findings can be interpreted as follows: The reliability estimate and separation indices in Analysis 1 were inflated by the presence of test items that were somewhat redundant and therefore limited in the amount of unique information they contributed to the estimation of person ability. Therefore, Analysis 3 provides a more accurate portrayal of person reliability.

Score Correlations

Generally, in both the Rasch dichotomous and partial credit models, the order of Rasch person measures is identical to the order of raw scores. For this reason, it is perhaps unsurprising that the correlations between raw scores and Rasch measures were nearly perfect (Table 3). The reason for correlations less than 1.00 is that the relationship was slightly curvilinear (Figure 1).

Table 3 *Matrix of Pearson correlations for raw scores and Rasch person measures*

Scoring	Hu & Nation Raw Scores	Analysis	
		1	2
Analysis 1 (Rasch)	0.998 [.997, .999]		
Analysis 2 (Rasch)	0.997 [.995, .998]	0.999 [.999, 1.000]	
Analysis 3 (Rasch)	0.995 [.992, .997]	0.999 [.998, .999]	0.999 [.999, 1.000]

Note: The assumption of normality for the use of Pearson correlations was confirmed with the Kolmogorov–Smirnov test; Hu & Nation raw scores, D = .086, p = .738; Analysis 1, D = .094, p = .630; Analysis 2, D = .109, p = .439; Analysis 3, D = .109, p = .448.

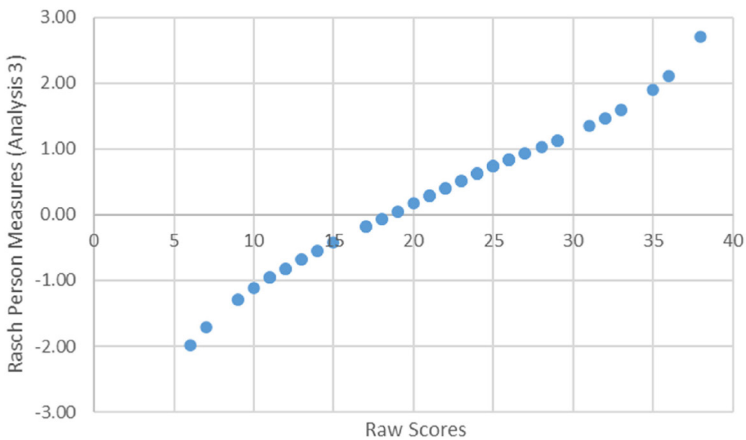

Figure 1 *Plot of Hu & Nation Raw Scores vs Rasch Person Measures from Analysis 3.*

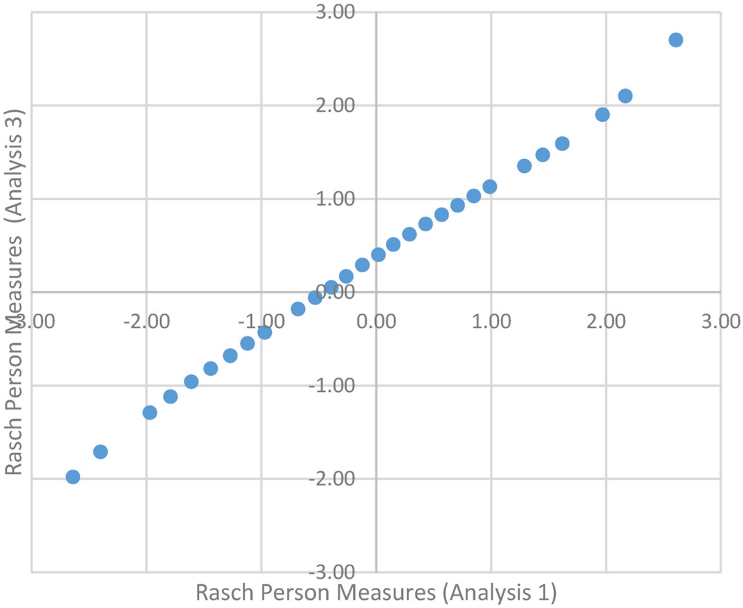

Figure 2 *Plot of Rasch Person Measures from Analysis 1 (LID present) and Analysis 3 (LID removed).*

Similarly, the correlation between Rasch measures in Analysis 1 (with LID present) and Analysis 3 (with LID removed) was nearly perfect (r = .999), again with slight curvilinearity present (Figure 2). This very strong relationship between person measures with and without LID differs from some of Marais and Andrich's (2008) findings, in which the correlations ranged from .80 to .98 depending on the strength of the residual correlations, the difficulty of the correlating items, and the pattern of residual correlations (across the entire test or within sub-sets of items). The difference between Marais and Andrich's work and the present study appears to be that whereas in the former, every item was locally dependent with at least one other, in the present study, despite using a test format likely to induce item dependency, many items were, in fact, statistically locally independent. Hence, the impact of LID on test reliability and score correlations was relatively minor.

Conclusions

This study examined Hu and Nation's (2000) cued-recall reading comprehension test for possible effects of local item dependency. Although LID was present, its impact on measurement was small and the interpretation of scores was virtually unchanged. LID inflated estimates of person reliability by .05, but when high residual correlations were completely removed by forming polytomous super-items, measurement consistency remained satisfactory for research purposes for this group of participants (Nunnally & Bernstein, 1994). It should be noted, however, that these participants represented learners of English as a foreign language from just three L1 groups, all in Asia. Although data from any future use of this test should again be evaluated on its own merits, and person reliability should be reported after the removal of high residual correlations, our findings suggest that LID is not particularly problematic in this test when the purpose of assessment is to separate learners according to the tested construct.

Acknowledgment

We would like to thank Chris Nicklin for his very helpful feedback and suggestions on an earlier draft of this paper.

References

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 18*(1), 6–40. https://doi.org/10.1177/0265532220927487

Baghaei, P. (2008). Local dependency and Rasch measures. *Rasch Measurement Transactions, 21*(3), 1105–1106. https://www.rasch.org/rmt/rmt213b.htm

Christensen, K. B., Makransky, G., & Horton, M. (2017). Critical values for Yen's Q_3: Identification of local dependence in the Rasch model using residual correlations. *Applied Psychological Measurement, 41*(3), 178–194. https://doi.org/10.1177/0146621616677520

Council of Europe (2001). *Common European framework of reference for languages.* Strasbourg: Council of Europe.

Fan, J., & Bond, T. (2019). Applying Rasch measurement in language assessment: Unidimensionality and local independence. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment* (Vol. I, pp. 83–102). Routledge.

Hu, M., & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*, 403–430. http://scholarspace.manoa.hawaii.edu/bitstream/10125/66973/1/13_1_10125_66973_rfl131hsuehchao.pdf

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Laufer, B., & Cobb, T. (2020). How much knowledge of derived words is needed for reading? *Applied Linguistics, 41*(6), 971–998. https://doi.org/10.1093/applin/amz051

Linacre, J. M. (2022). *Winsteps Rasch measurement*, Version 5.2.2 [Computer program]. https://winsteps.com

Marais, I., & Andrich, D. (2008). Effects of varying magnitude and patterns of response dependence. *Journal of Applied Measurement, 9*(2), 105–124.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149–174. https://doi.org/10.1007/bf02296272

McLean, S. (2021). The coverage comprehension model, its importance to pedagogy and research, and threats to the validity with which it is operationalized. *Reading in a Foreign Language, 33*(1), 126–140. https://nflrc.hawaii.edu/rfl/item/528

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8*, 125–145. https://doi.org/10.1177/014662168400800201

Appendix A

Hu and Nation's (2000) Cued-Recall Reading Test

Below is the reading test used in the present study. It is identical to that used by Hu and Nation (2000) except for the following:

1. In the reading, (a) *hullo* was changed to *hello*; (b) hyphens were replaced with spaces in *school-yard, front-door, letter-box, back-door, flower-bowl, telephone-operator, dining-room*, and *police-officer*; and (c) the first letter of *Governor* in the phrase *Governor of the hospital* was changed to lower case.
2. Small changes were made to the instructions for clarity, and they were translated to the learners' first languages.

Instructions

Read the passage. You can read it as many times as you like. There is no time limit. If it is helpful for you, you can mark up the text with your pencil or pen.

Afterwards, you will be asked to answer some questions and write down as much information as you could remember from the text without a time limit, but you will not be able to look back at the text. Then the second comprehension measure you will be asked to do is some multiple-choice questions. The results of those tests will only be used for our research.

Now, before we begin, please raise your hand if you have any questions.

The Escaped Madman

I was in the middle of making a cake for dinner when the telephone rang. I washed my hands and went to answer it.

"Hello," said an anxious voice when I had picked up the receiver, "could I speak to Mrs. Scott, please?"

"Speaking," I answered.

"Mrs. Rosalind Scott?" the voice went on, even more anxiously.

"Yes," I answered. "What's the matter?" By now I was getting a bit frightened myself. Thoughts of my husband being in a railway crash, or of one of the children breaking his leg playing football in the school yard raced through my mind.

"This is the Northfields Hospital," said the voice. "One of our patients has escaped, and one of his friends heard him say that he was going to kill you. We thought we'd better warn you."

"Oh – yes – thank you," I answered and stopped to think very quickly. "What's the name of this – er – patient?" I then went on.

There was a pause at the other end, as if someone was searching through a list. Then, "George Hitchcock is his name," came the reply. "He's short and dark-haired and very dangerous."

"But why does he want to – er – kill me? I've never even heard of him."

A strange sound came from the telephone and it went dead. Thoughts of people cutting telephone wires before attacking a house flashed through my brain. Would a madman think of doing a thing like that?

I was still holding the receiver when there was a knock at the front door. I dropped the receiver and raced to the door. I locked it just as the man outside was beginning to open the cover of the letter box. Then I rushed to the back door, knocking a small table over on my way and breaking the flower bowl on it to pieces. The water made the floor smooth and I slipped and crashed into a chair, knocking that over too. However, I reached the back door, locked it and then went around the house making sure all the windows were shut. They had no bars on them, so it would be easy to break one open.

Meanwhile, the telephone operator, finding I had not put my telephone down and hearing the crashes and knocking, thought there was a fight going on in my house and telephoned the police. In a few minutes, I heard heavy footsteps advancing up the drive and there was a firm sound on the knocker of the front door.

"Who's that?" I said in a trembling voice.

"Police," came the answer. "Is there anything wrong, Lady?"

I thought fast. "Come round to the window so that I can see you," I said. The heavy footsteps circled the garden, and two policemen – a police officer and a younger policeman – appeared at the dining room window. I ran joyfully to the front door to let them in. The mailman was standing a few yards off, looking at me in a way that struck me as unusual.

"Did you catch the madman?" I asked the police officer.

"What madman?" he asked, clearly completely confused.

"Why, the one that was knocking at my door. He escaped from the hospital, and his name's George Hitchcock. George Hitchcock."

"George Hitchcock?" said the officer, even more confused than before. "But that's the name of the governor of the hospital."

"And it was me that knocked at your door," said the mailman.

We looked at each other – the two policemen, the mailman and I. Fright, surprise and suspicion followed each other across our faces. Then the police officer said, "Somebody must have been having a joke with you, lady, and a stupid joke at that."

We all burst into laughter.

When you have finished carefully reading the passage as many times as you need to understand it well, please quietly raise your hand. Your teacher will bring you the URL and QR code for the reading comprehension questions.

Cued-Recall Questions

Try to retell the story by answering all of the following questions. Put as much information as you can.

1. Who is telling the story?
2. What was happening at the very beginning of the story?
3. What happened next?
4. What were her feelings when she spoke on the telephone?
5. Why did she feel this way?
6. What was she told on the telephone?
7. What happened next?
8. What did she think?
9. What happened at the door?
10. What did she do?
11. What were her feelings when she went through the house?
12. Why did she feel this way?
13. What did she do as she went through the house?
14. Who came to her house?
15. Why did they come?
16. What did she think when someone came to her house?
17. What did she tell them to do?
18. Why did she tell them to do this?
19. Who did she see?
20. What did she do when she saw them?
21. Who else did she see?
22. What was he doing and thinking?
23. What did she ask them?
24. What did they say?
25. How did each of them respond at first?
26. What was she told about after that?
27. What happened at the end?

Appendix B

Hu and Nation's (2000) Cued-Recall Reading Test Item List and Key

Item	Prompt	Key
1_1	Who is telling the story?	Mrs. Scott
1_2	Who is telling the story?	the housewife
1_3	Who is telling the story?	the author
2_1	What was happening at the very beginning of the story?	she was making dinner
2_2	What was happening at the very beginning of the story?	when the telephone rang
3_1	What happened next?	an anxious voice on the phone wanted to talk to her
4_1	What were her feelings when she spoke on the telephone?	she was frightened/alarmed/anxious at that time
5_1	Why did she feel this way?	because the anxious voice made her think that her husband or children might be in an accident
6_1	What was she told on the telephone?	one of the patients escaped from the hospital
6_2	What was she told on the telephone?	and wanted to kill her
6_3	What was she told on the telephone?	the patient's name is George Hitchcock
7_1	What happened next?	she wanted to know why he wanted to kill her
7_2	What happened next?	then the phone went dead
8_1	What did she think?	it made her think that the escaped man came to her house cutting telephone wires
9_1	What happened at the door?	there was a banging sound at the front door
9_2	What happened at the door?	a man outside was beginning to open her mailbox
10_1	What did she do?	she raced to the door
10_2	What did she do?	she locked the door
11_1	What were her feelings when she went through the house?	she was nervous and frightened
12_1	Why did she feel this way?	she thought the escaped man might come or break into her house to kill her
13_1	What did she do as she went through the house?	she made a mess knocking a small table
13_2	What did she do as she went through the house?	smashing the flower bowl
13_3	What did she do as she went through the house?	crashing into the chair
13_4	What did she do as she went through the house?	she reached the back door and locked it
13_5	What did she do as she went through the house?	and making sure all the windows were shut

Item	Prompt	Key
14_1	Who came to her house?	the two policemen
15_1	Why did they come?	the telephone operator heard the noise
15_2	Why did they come?	(the telephone operator) thought there was a fight
15_3	Why did they come?	(the telephone operator) telephoned the police
16_1	What did she think when someone came to her house?	she thought it might be the madman
17_1	What did she tell them to do?	she told them to come round to the window
18_1	Why did she tell them to do this?	she wanted to make sure that they were real policemen
19_1	Who did she see?	the policemen
20_1	What did she do when she saw them?	she ran joyfully to the door to let them in
21_1	Who else did she see?	the postman
22_1	What was he doing and thinking?	he was standing a few yards away
22_2	What was he doing and thinking?	he thought Mrs. Scott behaved quite unusual
23_1	What did she ask them?	she asked if the policemen caught the madmen
24_1	What did they say?	the policemen said that George Hitchcock was the governor
24_2	What did they say?	the postman told her it was him knocking at the door
25_1	How did each of them respond at first?	they looked at each other in surprise and suspicion
26_1	What was she told about after that?	someone must have played a joke with her
27_1	What happened at the end?	they all burst into laughter

Note: Horizontal lines in the table body denote the separation of items into super-items for Analysis 3.