A Reliability Analysis of Summative Cloze Test Formats

Trevor HOLSTER

Introduction

Bachman and Palmer (1996) argue for the centrality of usefulness in language testing, composed of reliability, construct validity, authenticity, interactiveness, impact, and practicality, in test design, raising the problem of incommensurable goods and conflicts between them, and practicality may become an overriding concern, leading to compromised validity. Weir (2005, p.12) reminds us that validity resides in test scores, not the test itself, and Bachman and Palmer (1996) explain validity in terms of inferences drawn from test scores, so a useful test allows valid inferences to be drawn regarding the purpose of the test. Thus test designers need to work from a clear specification of the ability about which inferences are desired, and the purposes of the inferences, so assessment tasks appropriate for one purpose may be unsuitable for another (Shohamy, 1992). Thus, teachers experienced only in formative assessment, may be unprepared for the rigorous procedures required for validity in summative testing. (Bachman, 1990; Hughes, 2003; Shohamy, 1992; Weir, 2005).

The construct in question will vary according to whether judgments are needed about proficiency, placement, achievement, or diagnosis, as Brown (J. D. Brown, 2005) explains. Proficiency tests are not specific to any curriculum, in contrast to placement tests, which aim to measure a narrowly specified set of abilities relevant to a specific curriculum. Both must be able to reliably compare individuals, so will be norm-referenced. An achievement test also aims to determine whether objectives have been met, while a diagnostic test aims to target remedial instruction, so these two types of tests are criterion-referenced to a specific curriculum. A general proficiency test might not adequately sample curriculum content, s placement, achievement, or diagnostic decisions may not be valid. As Hughes (2003, p.34) states, " Tests for which validity information is not available should be treated with caution. ", and thus it is incumbent on the test designer to make such information available to users. Westrick (2005) supports the view that proficiency tests may be invalid for placement, and recommends in-house tests, but given practicality concerns, the months that Shohamy (1992) considers necessary for test development and validation may not be available, so, in the author's experience, tests are often developed in timeframes inadequate to meet the minimal standards of validity and reliability that experts recommend (Bachman, 1990; Bachman & Palmer, 1996; H. Brown, 2004; J. D. Brown, 2005; Hughes, 2003; Spolsky, 1985; Weir, 2005). Given that the criticism of commercial tests on grounds of reliability and validity, in-house tests are defensible only if they improve the validity of decisions. Additionally, researchers such as Cook (2007) and Westrick (2005) argue for the im-

portance of placement tests on the grounds of improved instruction, in other words, on the grounds of test impact (Bachman & Palmer, 1996), and when placement decisions are irreversible, this gives makes them high-stakes, so validity and reliability are central concerns, following Westrick (2005).

Because placement, achievement, and diagnostic tests are curriculum specific, the curriculum and tests must be integrated, and all classes must focus on the same mechanisms of learning, so placement decisions are invalidated if teachers deviate from the institution wide approach. Therefore decisions regarding test design must begin with defining objectives (Hughes, 2003), necessitating a detailed top-down course specification, requiring a long-term curriculum and test development program, which is time consuming.

A further difficulty concerns course grades. Inoue (2006) describes curriculum reform in a Japanese public university, requiring inter-teacher consistency in course grades, implying some degree of norm-referencing to allow calibration points for the grade levels. If classes are not streamed, teachers should encounter a representative sample of learners, allowing norm-referencing, but this is not possible with streamed classes, leading to questions of fairness and arguments for standardized testing or item banking (Bachman, 2004; Henning, 1987).

Item banking allows the linguistic and statistical properties of items to be analyzed and matched to course specifications, allowing reliable inter-candidate comparisons. The key to this is equating test difficulty through " anchoring " (Bond & Fox, 2007; Henning, 1987), based on latent trait analysis, allowing the " banking " of items of known difficulty, based on an initial set of reliable anchor items. Reliability is essential because, unless a test consistently returns the same scores, it cannot be measuring a stable construct, precluding valid inferences. Therefore, as part of ongoing research funded by a public university, the usefulness of cloze tests as anchor items was investigated.

The cloze format was selected due its integrative nature, excellent construct validity, and practicality (H. Brown, 2004; Hughes, 2003) allowing evaluation of large numbers of items. Classroom teachers can easily construct cloze test items, calibrate them against the anchor items, and reliably test large groups, in contrast to the exhausting demands of direct open-ended performance tests, such as speaking proficiency or essay writing (Hughes, 2003).

Method and Results

Two cloze listening tests were administered to approximately 170 learners at two Japanese public universities as part of end of semester reviews to identify suitable items for a semester final test. Learners were told that the tests were being piloted and excessively difficult questions because would be eliminated from the final test. The consent form used earlier in the research project, in an attempt to ascertain the validity of the tests described by Cook (2007), was administered and, it was explained that participation was voluntary and unconnected to course grades. 112 learners agreed to participate. One test was based on homework dialogues from the textbook (Graham-Marr, 2007), with the publisher's permission. Tests were constructed by selected deletion of approximately every seventh word. In order to provide more difficult items, a second cloze test was constructed using dialogues from a reproducible IELTS resource book

(Brook-Hart, 2005) and administered twice. Additionally, a self-assessment questionnaire based on Cook (2007) was administered concurrently to complete the requirements of the university funded research. Two other tests were administered to sub-groups of learners, a 30 item multi-choice sentence completion test provided by a former colleague, and a combined 15 item multi-choice sentence completion and 20 item paragraph reordering test. These items were based on reproducible resource books by Brook-Hart (2005) and Altano (2005).

Initial analysis was based on classical test theory (CTT) notions of facility values and discrimination indices (Henning, 1987). The paragraph reordering items proved difficult to score, resolved by checking each sentence for correct ordering compared with its immediate neighbors. The CTT analysis suggested a high proportion of unsuitable items, and the need for a much longer test, consistent with Hughes (2003, p.44). Unfortunately, before the analysis was reported, a unilateral decision to reduce the number of items was made in the author's absence. The shortened version of the test could not possibly allow valid decisions, raising major ethical questions, and leading to the withdrawal of these items.

All the tests were analyzed under latent trait theory, using the Bond & FoxSteps software package (Bond & Fox, 2007), and item misfit was determined, as illustrated in Figure 1, showing the self-evaluation items. The horizontal axis shows item measure, scaled to 5 units per logit, while the vertical axis shows item misfit. We can see that many of these items do not measure the same trait as the overall test, given the normal limit of 2 logits misfit, so valid inferences cannot be drawn. By recursively removing misfitting items, following Bond and Fox's (2007) guidelines, 12 items displaying a shared trait were found, shown in Figure 2. The other tests were similarly analyzed, and internal reliability determined. Next the test-retest reliability of the IELTS based cloze was calculated, finding a significant and meaningful correlation of 0.992, indicating a coefficient of determination of 0.984, so only 1.6% of variance is not attributed to a shared ability, indicating excellent reliability.

Figure 1: Self Assessment Item Measure and Misfit

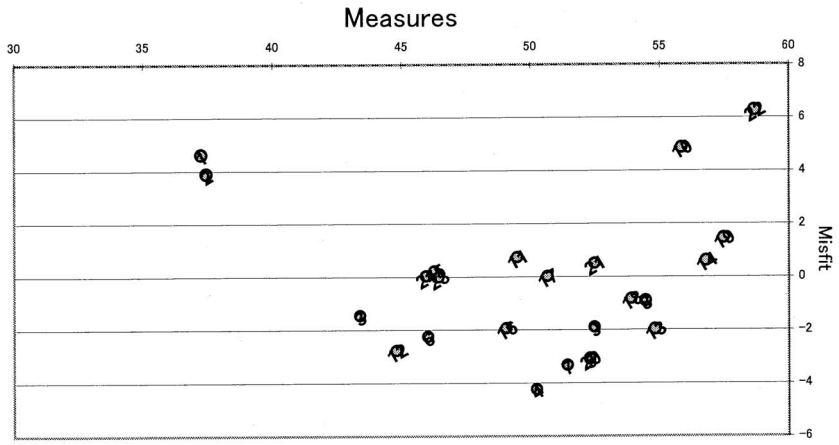


Table 1 shows internal reliability figures. The multi-choice writing screening test predictably fared very poorly, due to being too short. The longer test sections and the combined tests showed better reliability, with the combined cloze and self-assessment showing very high figures, as did the self-assessment. However, the likelihood of a halo effect means the self-assessment reliability must be treated with caution.

Table 2 shows the correlations between test sections. The cloze scores generally correlate significantly and meaningfully with both the grammar and Center test scores, with coefficients of determination of 0.48 and 0.45, meaning that 48% and 45% of score variance is shared, indicating a shared construct, supporting validity as a test of general proficiency. The self-assessment shows coefficients of determination of 0.077, 0.128, and 0.119 with the overall cloze test, Center test, and overall grammar test respectively, and thus no evidence of validity as a test of proficiency, raising doubts about the construct validity of the self-assessment, so valid inferences are not possible and it is incumbent on designers to demonstrate its validity before it can be used to make decisions that affect students.

Figure 2: Revised Self Assessment Item Measure and Misfit

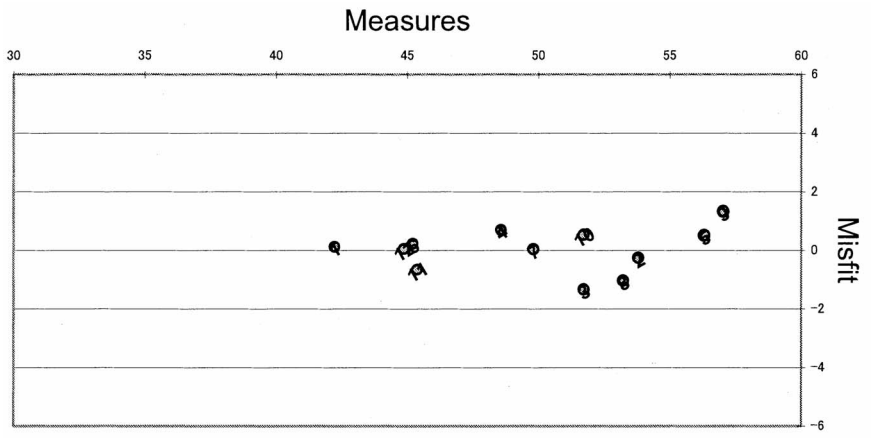


Table 1: Internal Reliability of Test Sections

Test	Section	Cronbach's Alpha	N of Items
Grammar	TOEFL MC	.217	15
	Anchor MC	.755	29
	All MC	.730	44
	Paragraphs	.771	18
	Complete	.813	63
Cloze	Course Text	.888	53
	IELTS	.898	74
	Complete	.941	123
Self Assessment		.913	12
Combined Test		.960	204

Figure 3 shows the textbook based cloze item difficulty, scaled to a mean of 50, with a mean ability of 61. This test might serve as an achievement or placement test, but lacks the range needed in a proficiency test as there are insufficient difficult items.

Table 2: Correlation Matrix for Test Sections

		Cloze Overall	Self Assessment	Center Test	Grammar
Cloze Overall	Pearson Correlation	1	.278(**)	.695(*)	.673(**)
	Sig. (2-tailed)		.003	.012	.000
	N	112	110	12	33
Cloze Text	Pearson Correlation	.924(**)	.217(*)	.572	.609(**)
	Sig. (2-tailed)	.000	.023	.052	.000
	N	112	110	12	33
Cloze IELTS	Pearson Correlation	.908(**)	.290(**)	.612(*)	.584(**)
	Sig. (2-tailed)	.000	.002	.034	.000
	N	112	110	12	33
Self Assessment	Pearson Correlation	.278(**)	1	-.359	.345
	Sig. (2-tailed)	.003		.252	.053
	N	110	110	12	32
Center Test	Pearson Correlation	.695(*)	-.359	1	n.a.
	Sig. (2-tailed)	.012	.252		
	N	12	12	12	2
Grammar Overall	Pearson Correlation	.673(**)	.345	-1.000(**)	1
	Sig. (2-tailed)	.000	.053	.	
	N	33	32	2	33

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Figure 3: Item Map for Cloze Textbook Test

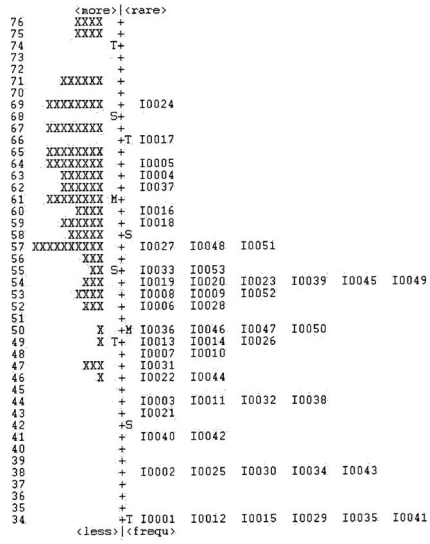


Figure 4 shows the IELTS based cloze test, with a difficulty range exceeding the ability range, and a standard deviation of 11. Mean ability has dropped to 54, with a standard deviation of 6. This test could function as a measure of general ability.

Figure 4: Item Map for Cloze IELTS Test

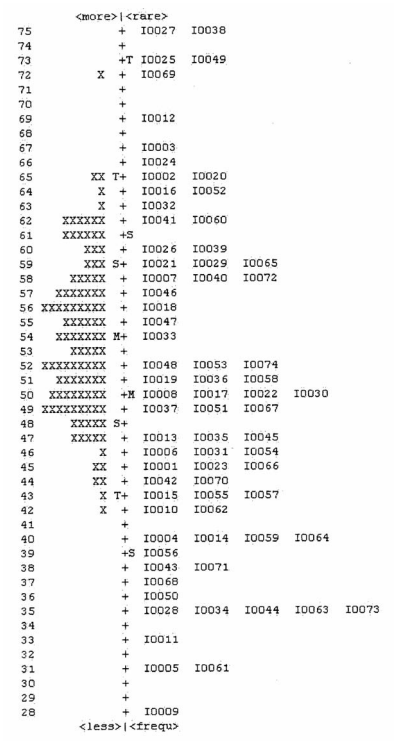
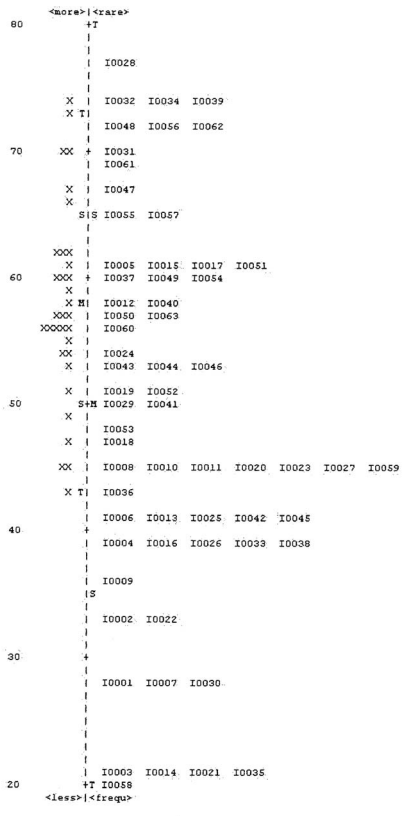


Fig. 5, the item map for the combined grammar test, shows insufficient difficult items, reflecting the purpose of determining cut points. Valid inferences for placement or diagnostic purposes might be possible, but a greater range of difficulty and considerably more items are needed. Reducing this test to less than 20 items would have been extremely ill advised.

Figure 6 shows the self-assessment item map, with an insufficient range and distribution of item difficulty, so discrimination of candidates for placement purposes is not possible. All responses were rescaled to a common scale then analyzed as a combined test to determine suitability for use as a combined test battery. Figure 7 shows the item measure and misfit, and we can see the self-assessment items are extremely over-fitting, suggesting a pronounced halo effect. By removing misfitting items, an acceptably fitting set of items was found, composed of 183 of the original 202 non-self-assessment items, with a range comfortably exceeding the ability of the target group. The resulting measure and fit chart, shown in Figure 8, describes a group of items with an internal reliability (Cronbach's alpha) of .956, giving a good provisional list of anchor items to begin construction of an item bank.

Figure 5: Item Map for Combined Grammar Test



Determining validity for diagnostic purposes would require a full factor analysis, but one immediate question is to what degree vocabulary is tested, so the 70 easiest and 70 most difficult

cloze test answers were compared using VocabProfile (Cobb, 2007; Heatley & Nation, 1994). The results, given in Table 3, show differences between the two lists. 53% of the easiest items are 1K level content words, compared with only 38% of the most difficult items. Approximately 18% of the most difficult items were from the 1K-2K list, compared with 8% of the easiest answers. The significance of this was not calculated, but the raw data supports the hypothesis that vocabulary knowledge is a major component of the construct tested and that further investigation is warranted.

Figure 6: Item Map for Self Assessment.

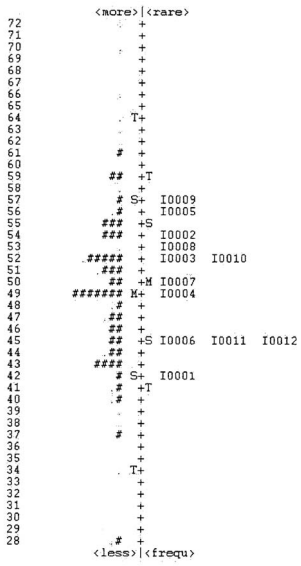


Figure 7: Combined Test Item Measure and Misfit.

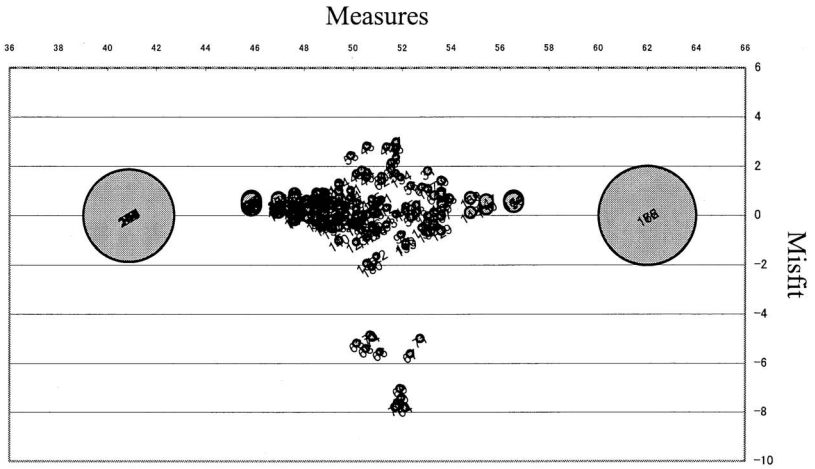


Figure 8: Revised Combined Test Item Measure and Misfit.

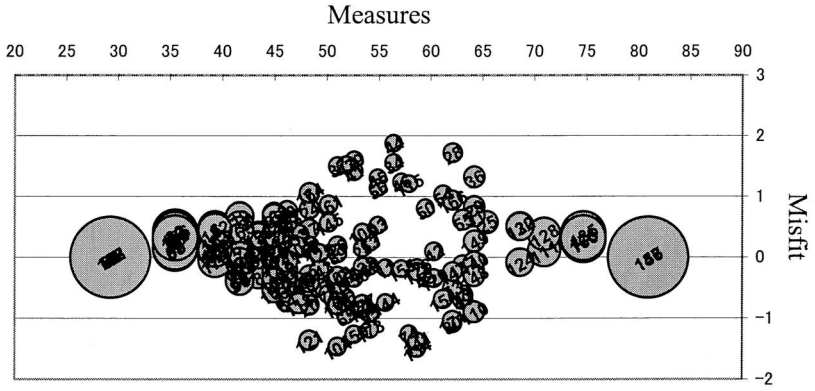


Table 3: Vocabulary Distribution by Difficulty

	Percent		Families		Types		Tokens	
	Bottom 70	Top 70	Bottom 70	Top 70	Bottom 70	Top 70	Bottom 70	Top 70
K1 Words:	85.53%	72.37%	49	48	53	50	65	55
Function:	32.89%	34.21%					25	26
Content:	52.63%	38.16%					40	29
K2 Words:	7.89%	18.42%	4	12	4	12	6	14
1k+2k	93.42%	90.79%						
AWL:	2.63%	1.32%	2	1	2	1	2	1
Off-List:	3.95%	7.89%			3	5	3	6
	100%	100%	55+?	61+?	62	68	76	76

Conclusions

The major objective of this study was to investigate the reliability of cloze listening tests to identify suitable anchor items for an item bank, and to determine validity for general proficiency, achievement, placement, and diagnostic tests. The cloze tests showed exceptional reliability and the limited assessment of concurrent validity suggests they are good measures of general language ability, but a more sophisticated assessment of concurrent validity is needed. The analysis conducted here was designed to be simple enough that classroom teachers could replicate it and establish an item bank without specialist help. The cloze tests are very quick and easy to construct, so the primary objective of this study was successful.

A secondary objective was to investigate the usefulness of other test formats, in order to establish an item bank allowing construction of test batteries of different item types, but this was only partially successful. The paragraph reordering items proved difficult to grade, and may re-

quire machine grading using a computer algorithm if untrained graders are to be used. The multiple-choice items also showed that a longer test is required to achieve acceptable levels of usefulness, and are difficult and time-consuming to construct, moderate, and validate. However, the discrete point measurement that these items allow is essential for diagnostic purposes, so further development of these is warranted. This will require factor analysis to identify the constructs each item measures, so a large number of items and considerable time will be required to establish a useful bank of items, but it is feasible in the longer term with modest resources.

The analysis of the self-assessment found that half the items were misfitting, leading to serious doubts about construct validity. Once these items were removed, the 12 remaining items showed internal consistency, but an inadequate range of difficulty, and did not fit the construct assessed by the other assessments, casting doubt on whether they measure any construct relevant to language learning. It is highly likely that the internal reliability is due largely to a halo effect, not an underlying construct related to language ability. Until these issues are addressed by the test designers, this assessment cannot be considered suitable for comparing individual candidates, as Oscarson (1997) makes clear, so is not suitable for use as a proficiency, placement, achievement, or diagnostic test. It is therefore recommended that this assessment be discontinued.

The analysis of the cloze test scores relationship to vocabulary difficulty showed promising results, but a more sophisticated analysis of a larger data set will be needed before definitive results can be reported. This will be a focus of future studies, combined with analysis of other features contributing to item difficulty and further investigation of concurrent validity. Once completed this should allow construction of an item bank of well specified items necessary for diagnostic purposes.

References

Altano, B. (2005). *Testing academic reading processes.* Michigan: The University of Michigan Press.

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. (2004). *Statistical analyses for language teachers.* Cambridge: Cambridge University Press.

Bachman, L., & Palmer, A. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Bond, T., & Fox, C. (2007). *Applying the Rasch model* (2 ed.). London: Lawrence Erlbaum Associates.

Brook-Hart, G. (2005). *Instant IELTS.* Cambridge: Cambridge University Press.

Brown, H. (2004). *Language assessment: Principles and classroom practices.* White Plains: Longman.

Brown, J. D. (2005). *Testing in language programs.* New York: McGraw-Hill.

Cobb, T. (2007). Web Vocabprofile. Retrieved September 16, 2007, from http://www.lextutor.ca/vp/

Cook, M. (2007). *The place of placement tests* Paper presented at the Nagasaki JALT Annual My Share Session.

Graham-Marr, A. (2007). *Communication spotlight: High beginner.* Tokyo: Abax.

Heatley, A., & Nation, P. (1994). Range: Victoria University of Wellington.

Henning, G. (1987). *A guide to language testing.* Boston: Heinle & Heinle.

Hughes, A. (2003). *Testing for language teachers.* Cambridge: Cambridge University Press.

Inoue, N. (2006). What's going on inside the pine tower of babel: Foreign language curriculum reform in a Japanese university. *Languages and Cultures Series., 16,* 87-115.

Oscarson, M. (1997). Self assessment of foreign and second language proficiency. In C. Clapham & D. Corson (Eds.), *Encyclopedia of Language and Education* (Vol.7: Language testing and assessment, pp. 175-187).

Shohamy, E. (1992). New modes of assessment: the connection between testing and learning. In E. Shohamy & A. Walton (Eds.), *Language Assessment for Feedback: Testing and Other Strategies.* Dubuque: Kendall Hunt.

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing, 2*(1), 31-40.

Weir, C. (2005). *Language testing and validation.* New York: Palgrave Macmillan.

Westrick, P. (2005). Score reliability and placement testing. *JALT Journal, 27*(1), 71-92.