English Language Teaching; Vol. 14, No. 5; 2021
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

Identifying Guessing in English Language Tests via Rasch Fit Statistics: An Exploratory Study

David Coniam[1], Tony Lee[1] & Leda Lampropoulou[1]

[1] LanguageCert, PeopleCert, 3 Korai Street, Athens 10564, Greece

Correspondence: David Coniam, PeopleCert, 3 Korai Street, Athens 10564, Greece.

Received: January 5, 2021 Accepted: March 29, 2021 Online Published: April 14, 2021

doi: 10.5539/elt.v14n5p23 URL: https://doi.org/10.5539/elt.v14n5p23

Abstract

This article explores the issue of identifying guessers – with a specific focus on multiple-choice tests. Guessing has long been considered a problem due to the fact that it compromises validity. A test taker scoring higher than they should through guessing does not provide a picture of their actual ability. After an initial description of issues associated with guessing, the article then outlines approaches which have been taken to either discourage test takers from guessing or which attempt statistically to handle the problem. From this, the article moves to a novel way of identifying potential guessers: from the post hoc use of Rasch fit statistics. Two datasets, each consisting of approximately 200 beginner level English language test takers were split into two. In each dataset, half the test takers' answers were randomised – to approximate guessing. Results obtained via a Rasch analysis of the data was then passed to an analyst who used the Rasch fit statistics to identify possible guessers. On each dataset, 80% of guessers were identified.

Keywords: guessing, Rasch, fit statistics, English language

1. Introduction

The key concept in assessment is generally considered to be *validity*. Validity (see, e.g., Messick, 1989; Bachman & Palmer, 2010) may be framed as constituting the extent to which a given test score can be interpreted as an indicator of the key abilities or constructs being measured. There are a number of issues which may be construed as "construct irrelevant variance" (Downing, 2002): cheating, testwiseness, teaching to the test, flawed test design, to name but a few. One construct irrelevant variance (CIV) that many have long grappled with, especially in multiple-choice (MC) tests is that of guessing. Guessing essentially increases measurement error in that it raises the possibility of correct test taker responses, and hence compromises validity.

Two issues that are interconnected relate to what can be done in terms of: one, identifying guessing; two, dealing with guessing.

While the focus of the current paper is on the first issue, it will be necessary to put both issues into perspective to give a sense of direction to the current paper. In a seminal paper on the effect of random guessing on test validity, Lord, some 50 years ago commented on the reality that while test takers may be advised not to guess, guessing on the part of test takers cannot really be militated against.

2. Identifying Guessing

While a considerable amount of effort has been expended into dealing with possible guessing, the literature on identifying guessing is surprisingly thin. Lord and Novick (1968) described an item's discriminating power as its effectiveness in discriminating among higher and lower achievers, and stated that the correlation between an item score and the overall test score (i.e., an item-total correlation) provides "a rough index of item discriminating power" (p. 331).

In a somewhat similar vein, Downing (2002) examined the issue of construct-irrelevant variance from the perspective of what he termed "flawed test questions". In his study, in which experts identified flawed items in a science test, the discrimination indices of the items and the overall KR20 reliability figure and the passing scores were lower for the flawed items than for well-constructed items. While Downing concluded with the observation that the use of flawed test questions may well have a negative impact on student performance, the study is a useful indicator of how statistics may have a role to play in identifying CIV issues, with guessing being one.

Boone & Staver, J. R. (2020) illustrate how the point measure correlation – a statistic provided by many Rasch analysis programs – may be used to identify possible errors in answer keys. While this may not be guessing per se, the construct-irrelevant variance that a flawed key brings to an analysis verges in a similar direction.

The work of Attali and Bar-Hillel (2006) focuses on the fact that with many MC tests, the correct answers will tend to be in the "middle positions", with test takers who guess then tending to select the middle options as their guesses.

A recent approach to identifying guessing – in the context of computer-based tests – has been that of response time – i.e., the time taken to record a response. The work of Wise (e.g., 2017, 2019) has illustrated how rapid-guessing (in computer-based tests) may be identified, and in part therefore dealt with. Certain studies have, however, found that the correctness of rapid guesses exhibits little relationship to overall test performance (Goldhammer et al., 2016).

3. Handling Guessing

Since guessing has long been an issue which has been seen to threaten test validity, methods of dealing with, or discouraging, guessing have been approached from a number of angles. One major approach which has morphed through various approaches – with its supporters and opponents – has been that of "formula scoring". Formula scoring involves penalising test takers for incorrect answers, deducting marks, thus attempting to get test takers to only attempt questions they feel sure of answering correctly. However, certain researchers (for example, Lord, 1964) have recommended that all test takers answer all questions – in effect forcing an equal willingness to guess on all test takers.

MC tests have generally been scored using a conventional number-correct scoring method (Kurz, 1999; Bereby-Meyer et al., 2002) where a test taker gets the score for the number of questions they have answered correctly, irrespective of whether they have guessed.

Kurz (1999) presents an overview of some of the different scoring formulas used to correct for guessing. In this method of negative marking, test takers are penalised for incorrect responses.

Another method proposed by Traub et al. (1969) rewards a student for not guessing – by awarding points for omitted items rather than penalising for incorrect responses.

Since it might be expected that in a four-option MC item, test takers will get 25% correct and where; in a 3-option MC item, test takers will get 33% correct, one proposal has been that the passing score should be raised to reflect this effect (see e.g., Lesage et al., 2013).

In summary, however, as Lesage et al. (2013) point out, in the 50 years that guessing on MC tests has been researched, evidence as to which method is the most robust for dealing with guessing is still lacking.

Guessing has also long been a concern in English language tests, having been explored and discussed in a wide range of assessment situations and contexts – from vocabulary to reading; see e.g., Haynes, 1984; Huibregtse et al., 2002; Vanhove & Berthele, 2015; Gyllstad et al., 2015.

One statistical model for dealing with guessing is the three-parameter IRT model (see e.g., Birnbaum, 1968; Waller, 1989). Using this model, it is possible to make corrections on the basis of correct answers. As LeBeau & McVay (2017) point out, however, large sample sizes at times inhibit the use of the model: Hulin et al. (1982) suggest a minimum of 1,000 subjects for accurate measurement in the 3PL model. While such samples may be accessible in large-scale tests, they will not be available to teachers in schools who may well be looking at class or, at most, school-size groups, and will be limited to a sample of a couple hundred students at most.

4. Background and Methods

Given that reducing guessing increases test validity (see Kurz, 1999), the question is therefore: To what extent can guessing, and in particular wild guessing, be identified?

The current study follows the line of argument regarding the use of statistics. In particular, the working hypothesis in the current study centres around the concept of model fit.

4.1 Model Fit

All measurements have expected outcomes: the measurement of a straight line requires, for example, that the object being measured has straight line edges. The one-parameter Rasch model, as a measurement model, expects assessment elements (persons and items) to conform to certain assessment properties in the model. Against this backdrop, the extent to which the assessment properties are adhered to by the assessment elements

illustrate the concept of 'model fit' and how this is articulated through what might be termed *broad* and *more focused* criteria.

Broad criteria are the *Point Measure* correlation, and *Infit* and *Outfit* mean square statistics (i.e., estimates of population variance, or standard error). A more focused criterion involves *Standardised Infit* and *Outfit* (i.e., Z-score) statistics. These statistics are outlined briefly below.

4.2 Point Measure Correlation

The point measure correlation (PTME) in the Rasch model is comparable to the conventional point biserial correlation. Negative PTME values indicate a lack of model fit.

4.3 Infit

Infit may be seen as the 'big picture' in that it scrutinises the internal structure of an item or person. High infit mean square values indicate rather scattered information within the item or person, providing a confused picture about the placement of the item or person. Very small infit values indicate only very small variation and, provide therefore, little information to articulate clear and meaningful judgments about an item or person.

4.4 Outfit

Outfit gives a picture of 'outliers', that is responses from persons or items that appear to be considerably out of line with where a person or item would expect to be placed. High outfit mean square values would flag an item or person as being out of line with the rest in the pool – hence an 'outlier'.

4.5 Standardised Z-Scores

The standardised Z-score for infit and outfit is a more refined model fit criterion, and an extension of the interpretation of mean square values. This is a t-test exploring how well the data fit the model.

4.6 Expected Values

As alluded to above, the central concept in Rasch is that of the 'fit' of the data to the Rasch model; i.e., the extent to which obtained values match expected values. A 'perfect' fit of the data to the model may be interpreted from three perspectives.

- good point measure correlations
- outfit and infit mean squares of 1.0
- standardised Z-scores of 0.0

Such figures would indicate that obtained values exactly match expected values.

Mean square values less than 1.0 generally indicate that observations are too predictable. In contrast, mean squares values above 1.0 indicate over-dispersion, the possibility of pure guessing; over-fit indicates near uniformity in response, the possibility of giving invariant answers – which can be construed as another way of guessing. While acceptable ranges of tolerance for fit vary, acceptable ranges are generally taken as from 0.7 (30% below expectations) to 1.3 (30% above expectations) (see Linacre, n.d. (a)).

Regarding standardised Z-scores, figures above 2.0 indicate considerable distortion or degradation in the measurement system (Linacre, n.d. (b)).

4.7 Data and Analysis

The data in the current study was drawn from tests produced by the international language assessment organisation *LanguageCert*. *LanguageCert* produce and administer a suite of tests – the International ESOL suite – which are aligned to the six CEFR levels: Preliminary (A1), Access (A2), Achiever (B1), Communicator (B2), Expert (C1) and Mastery (C2). The examination specifications reflect the requirements of the CEFR; test materials writers employ the highest international standards and have extensive expertise in, and knowledge and understanding of, the CEFR (Note 1).

Two datasets were constructed, both with beginner-level test takers. This level of ability was selected because test takers at this level are beginners, have a more restricted grasp of English and therefore are possibly more prone to guess. It should be noted that in the tests they were administered, there is no negative marking formula: no marks are deducted for incorrect answers.

The first dataset comprised a sample of 203 test takers who been graded at A1 level in terms of their English language standard under the CEFR. The second comprised a sample of 287 test takers who been graded at A2 level.

The datasets were first split into two, with the two groups in each dataset identified by their infit and outfit mean squares values.

Approximately half of the test takers' scores were left untouched on the basis of good fit to the Rasch model of the '1' threshold: all test takers' infit and outfit scores were between 0.95 – 1.05; that is, that the mismatch between expected-to-observed scores was only minimally (5% either way) under- or over-estimated. This group was labelled '1' – Unchanged.

The responses for the other half of the test takers were then randomised – to simulate random guessing. This group was labelled '2' – Randomised.

On t-tests run with each set of tests, the two groups were reported to be of equal ability, with no significance reported in group mean scores.

The data was then run through the software program *Winsteps* (Linacre, 2020), following which the results were passed to an analyst who attempted to identify the guessers, using the infit and outfit mean square output.

5. Hypothesis

The hypothesis in the current study is that – on the basis of high outfit or infit, or high standardised Z-score statistics – 80% of test takers may be identified as guessers.

6. Results

An analyst was then given test takers' mean square outputs; they were not given access to test takers' scores – which would have given the analyst hints and hence invalidated the exercise.

The following guidelines were given to the analyst in terms of coming to a decision about guessing:

1. Investigate negative person point measure correlations.
2. Investigate outfit, then infit.
3. Investigate mean squares, then standardised Z-scores.
4. Investigate high values, then low or negative values.

The analyst, in their examination of the mean squares, was instructed to assign one of three labels to each test taker. These, as mentioned, were: '1' – Not a Guesser, '2' – Definite Guesser, '3' – Unsure.

Cohen's kappa was used to report on coder agreement. According to McHugh (2012), a level of 0.6 for kappa indicates 'moderate' and a level of 0.8 or better 'strong' agreement.

Descriptive analyses are presented below, separately for each test. These are then followed by the crosstabulation picture, along with the figure for kappa. Tables 1 and 2 first present the results for Test A1.

6.1 Test A1

Table 1 presents the labels assigned in the original Test 1A dataset, with Columns 2 and 5 showing the number of cases in the different categories.

Table 1. Category descriptives – Test A1 (N=203)

Original	*Cases*	*Analyst*	*Cases*
Unchanged	107 (52.7%)	Not a Guesser	97 (47.8%)
Randomised	96 (47.3%)	Definite Guesser	87 (42.9%)
		Unsure	19 (9.3%)

As can be seen, the analyst suggested that 97 (47.8%) of the dataset were not guessers and 87 (42.9%) definite guessers. They were unsure about 19 test takers.

Table 2 presents a crosstabulation of the two sets of data, with Kappa calculated for inter-category agreement. The key cell showing agreements are in bold font.

Table 2. Category crosstabs – Test A1

		ORIGINAL		*Totals*
		Unchanged	Randomised	
ANALYST	Not a Guesser	**97 (90.1%)**	0 (0.0%)	97
	Definite Guesser	3 (2.8%)	**84 (87.5%)**	87
	Unsure	7 (7.2%)	12 (12.5%)	19
	Totals	107	96	203

In the dataset of 107 original unchanged answers, the analyst labelled 97/107 (90.1%) as 'non-guessers'. Of the 96 randomised answers, they labelled 84/96 (86.6%) as 'guessers'.

Kappa for inter-category agreement – that is, between the original categories and the analyst's verdicts – was 0.81 (p<.000) – 'strong' agreement in Hughes' (2012) terms.

Tables 3 and 4 present the results for Test A2.

The hypothesis in this case was, therefore, proven.

6.2 Test A2

Table 3. Category descriptives – Test A2 (N=287)

Original	*Counts*	*Analyst*	*Counts*
Unchanged	135 (47.0%)	Not a Guesser	147 (51.2%)
Randomised	152 (53.0%)	Definite Guesser	116 (40.4%)
		Unsure	24 (8.4%)

The analyst labelled 116 of the 287 subjects (40.4%) as definite guessers, and 147 (51.2%) as not guessers. They were unsure about 24 test takers.

Table 4 presents a crosstabulation of the two sets of data, with Kappa calculated for inter-category agreement.

Table 4. Category crosstabs – Test A2

		ORIGINAL		Totals
		Unchanged	Randomised	Total
ANALYST	Not a Guesser	**115 (85.2%)**	32 (25.6%)	147
	Definite Guesser	7 (5.2%)	**109 (71.2%)**	124
	Unsure	13 (9.6%)	11 (7.2%)	24
	Totals	135	152	287

In the dataset of 135 original unchanged answers, the analyst labelled 115/135 (85.2%) as 'non-guessers'. Of the 152 randomised answers, they labelled 109/152 (71.2%) as 'guessers'.

Kappa for inter-category agreement – that is, between the original categories and the analyst's verdicts – was, for this test, lower – 0.59 (p<.000), i.e., 'moderate' agreement in Hughes' (2012) terms.

The hypothesis in this case was, therefore, not proven.

7. Discussion

This study was exploring how guessers might be identified via the use of fit statistics produced via one-parameter Rasch analysis. The hypothesis was that 80% of test takers who exhibited high infit and outfit statistics might be identified as guessers. This hypothesis was accepted for Test A1, where 87.5% of random guessers were identified. This hypothesis was not accepted for Test A2, however, where only 71.2% of random guessers were identified – and where the 80% target threshold was not achieved.

While the study is limited in its scope in that it only involved two groups of beginner-level English language test takers, the methodology does illustrate the potential for identifying guessers. The one-parameter Rasch model does not require a very large sample, unlike its three-parameter cousin. The methodology in the current study has limited itself to low-ability test takers who might be more inclined to guess than more able ones; this is an issue that will need to be explored in further studies with other higher ability groups. It may be the case that the hypothesis was accepted for the absolute beginner test sample A1, because absolute beginners, having very little language to start with may be forced to guess more than learners with a higher proficiency of language, who may be making more 'judicious' attempts at items.

The value of the current study is in its relevance to small scale studies, where, for example, a teacher needs to identify guessers in a school exam to consider remedial work for students who may be experiencing difficulties. (The work of Ho et al. (2012) in using Rasch measurement with Hong Kong teachers is instructive here) Another area of possible use in where teachers are pretesting exam material, and only want to include answers by 'bona fide' respondents. Being able to eliminate certain guessers would enable pretest results to be seen to have greater validity. This is the key contribution of the current study: that reducing the amount of guessing in a test makes for getter validity in terms of how the scores may be interpreted.

References

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement, 40*(2), 109-128. https://doi.org/10.1111/j.1745-3984.2003.tb01099.x

Bachman, L., & Palmer, A. (2010). *Language Assessment in Practice*. Oxford, UK: Oxford University Press.

Bereby-Meyer, Y., Meyer, Y., & Flascher, O. M. (2002). Prospect theory analysis of guessing in multiple choice tests. *Journal of Behavioral Decision Making, 15*, 313-327. https://doi.org/10.1002/bdm.417

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: fundamental measurement in the human sciences* (2nd ed.). Mahwah, N.J.: Erlbaum.

Boone, W. J., & Staver, J. R. (2020). *Advances in Rasch Analyses in the Human Sciences.* Springer: Cham, Switzerland. https://doi.org/10.1007/978-3-030-43420-5

Downing, S. M. (2002). Construct-irrelevant Variance and Flawed Test Questions. *Academic Medicine, 77*(10), 103-104. https://doi.org/10.1097/00001888-200210001-00032

Fleiss, J. (1981). *Statistical Methods for Rates and Proportions*. Wiley VCH, New York.

Gyllstad, H., Vilkaitė, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL-International Journal of Applied Linguistics, 166*(2), 278-306. https://doi.org/10.1075/itl.166.2.04gyl

Haynes, M. (1984). Patterns and perils of guessing in second language reading. *TESOL, 83*, 163-176.

Ho, C. M., Leung, A. W. C., Mok, M. M. C., & Cheung P. T. M. (2012). Informing Learning and Teaching Using Feedback from Assessment Data: Hong Kong Teachers' Attitudes Towards Rasch Measurement. In Mok M. (Ed.), S*elf-directed Learning Oriented Assessments in the Asia-Pacific* (pp. 311-334). Education in the Asia-Pacific Region: Issues, Concerns and Prospects, vol 18. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-4507-0_17

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language testing, 19*(3), 227-245. https://doi.org/10.1191/0265532202lt229oa

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two-and three-parameter logistic item characteristic curves: a monte carlo study. *Applied Psychological Measurement, 6*(3), 249-260. https://doi.org/10.1177/014662168200600301

Jiang, S., Wang, C., & Weiss, D. J. (2016). Sample Size Requirements for Estimation of Item Parameters in the Multidimensional Graded Response Model. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.00109

Kurz, T. B. (1999). A review of scoring algorithms for multiple-choice tests. Paper presented at *the annual meeting of the Southwest Educational Research Association*, San Antonio, TX.

LeBeau, B., & McVay, A. (2017). *Validity of the three parameter logistic item response theory model for field test data*. ITP Research Series: University of Iowa. Retrieved from https://itp.education.uiowa.edu/ia/documents/Validity-of-the-Three-Parameter-Item-Response-Theory-Model-for-Field-Test.pdf

Lesage, E., Valcke, M., & Sabbe, E. (2013). Scoring methods for multiple choice assessment in higher education – Is it still a matter of number right scoring or negative marking? *Studies in Educational Evaluation, 39*(3), 188-193. https://doi.org/10.1016/j.stueduc.2013.07.001

Linacre, J. M. (2020). *Winsteps® Rasch measurement computer program*. Beaverton, Oregon: Winsteps.com

Linacre, J. M. (n.d. (a)). *Rasch power analysis: Size vs. significance: Infit and Outfit mean-square and standardized Chi-Square fit statistics*. Retrieved from https://www.rasch.org/rmt/rmt171n.htm

Linacre, J. M. (n.d. (b)). *Reasonable mean-square fit values*. Retrieved from https://www.rasch.org/rmt/rmt83b.htm

Lord, F. M. (1964). The effect of random guessing on test validity. *Educational and psychological measurement, XXIV*(4), 745-747. https://doi.org/10.1177/001316446402400401

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica, 22*(3), 276-282. https://doi.org/10.11613/BM.2012.031

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Rogers, H. J. (1999). Guessing in multiple choice tests. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 235-243). Amsterdam: Pergamon. https://doi.org/10.1016/B978-008043348-6/50019-X

Tate, R. (2003). A comparison of selected empirical methods for assessing the structure of response to test items. *Applied Psychological Measurement, 27*(3), 159-203. https://doi.org/10.1177/0146621603027003001

Vanhove, J., & Berthele, R. (2015). The lifespan development of cognate guessing skills in an unknown related language. *International Review of Applied Linguistics in Language Teaching, 53*(1), 1-38. https://doi.org/10.1515/iral-2015-0001

Waller, M. I. (1989). Modeling guessing behavior: A comparison of two IRT models. *Applied Psychological Measurement, 13*(3), 233-243. https://doi.org/10.1177/014662168901300302

Wise, S. L. (2019). An Information-Based Approach to Identifying Rapid-Guessing Thresholds. *Applied Measurement in Education, 32*(4), 325-336. https://doi.org/10.1080/08957347.2019.1660350

Note

Note 1. The CEFR has, over the past two decades, come to be accepted across Europe (and indeed beyond, with many countries linking their language curricula, syllabuses and examinations to the CEFR) as a specification of common standards across many different European languages. The CEFR lays out a set of common standards which permit employers and educational institutions to evaluate the language qualifications of test takers applying for employment or admission to education.