Chapter 17
Fit of Responses to the Model IV—Guessing

In this chapter, we study misfit due to guessing on multiple-choice items. The 3P model is discussed in *Chap. 18*: *Other Models of Modern Test Theory for Dichotomous Responses*.

Multiple-choice items are widespread in educational tests of proficiency. Guessing can be a threat to measurement, even on a well-constructed multiple-choice test. A person who does not know the correct answer either guesses randomly among *all* response alternatives, or based on partial knowledge first eliminates one or more of the alternatives and then selects randomly from the remainder.

Multiple-choice items are generally scored dichotomously and often analysed according to the dichotomous Rasch model. However, although the Rasch model has the desirable property of invariance realized through sufficiency, it makes no provision for guessing behaviour. From the Rasch paradigm perspective, the model is an operational rendition of fundamental measurement (Andrich, 2004) and the occurrence of random guessing is not a desirable property of a measurement system. So guessing is not a property of the model but of the data. When there is misfit between the data and the model, it is seen as an anomaly revealed in the data. If possible, new data should be generated that better conform to the model. This can be done in various ways, for example by improving the targeting of the test or changing test instructions. The ICC in Fig. 17.1 shows an item on which low proficiency persons guessed.

This item does not fit the Rasch model. When a model, like the simple dichotomous Rasch model does not fit the data, analysts in the traditional paradigm choose a more complex model, like Birnbaum's (1968) three-parameter (3P) model, on the grounds that it accounts better for the data (Andrich, 2004). The 3P, which models guessing in addition to different discrimination powers of items, is thought to more truly represent the behaviour of empirical items. In the 3P model the probability of a correct response is expressed as

$$\Pr\{X_{ni} = 1\} = c_i + (1 - c_i)P$$

D. Andrich and I. Marais, *A Course in Rasch Measurement Theory*,
Springer Texts in Education, https://doi.org/10.1007/978-981-13-7496-8_17

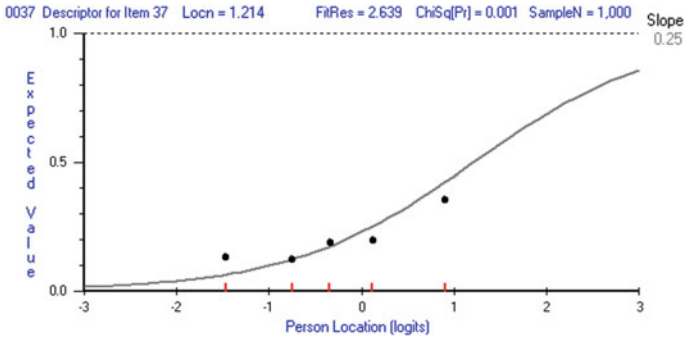

Fig. 17.1 ICC of an item with guessing

where $P = [\exp(\alpha_i(\beta_n - \delta_i))]/[1 + \exp(\beta_n - \delta_i)]$, α_i is the discrimination of item i and c_i is taken as the guessing parameter of item i. If it is not estimated but defined by the number of alternatives, then $c_i = 1/C$ where C is the number of response alternatives, which is the minimum probability that person n answers item i correctly.

Conceptual problems with the 3P model have been identified (e.g. Waller, 1973). In the 3P model guessing is considered an item parameter and, if not defined a priori, it is estimated along with other item parameters. However, it is persons rather than items who guess. The c parameter affects the probability of a correct response by *every* person to *every* item, and the model assumes all persons employ random guessing and they only guess on items to which they do not know the answer.

Guessing on a multiple-choice item occurs when a person does not know the correct answer, and this is more likely to be when a person has low proficiency relative to the item's difficulty (Andrich, Marais, & Humphry, 2012). Therefore lower proficiency students answer items correctly at a greater rate than they would if only their proficiency, and no guessing, played a role. An example of health outcomes where the symptoms in the data were similar to guessing was when older people were being assessed for memory functioning. They were given a list of words, and a short time later they were asked to recall each word. If the word was recalled, it was given a positive response. However, if they did not recall a word, they were given a prompt. Then if they recalled the word with a prompt, it was given a positive response. Thus persons who recalled a word with no prompt were given the same score of 1 as those persons who recalled a word with a prompt. Can you see why the effect here is similar to that of guessing?

Tailored Analysis

Waller (1973) proposed a procedure that *removes* the effect of guessing in estimating item and person parameters in the two-parameter (2P) model, another item response theory model which is discussed in Chap. 18. His 'Ability Removing Random Guess-

ing' (ARRG) procedure (Waller, 1973, 1989; Andrich et al., 2012) is based on the idea that guessing occurs when the item is too difficult for a person. Formally, guessing occurs when the probability of answering an item correctly is lower than the probability of answering the item correctly by chance. For example, in the case of an item with 4 response alternatives, these are the responses for a person where $\Pr\{X_{ni} = 1\} < 0.25$. To remove the effect of such guessing, all responses for which $\Pr\{X_{ni} = 1\} < 0.25$ are removed. This is in effect removing all the responses on items that are too hard for the person, a form of post-hoc tailored testing. Waller stressed that, rather than estimating and correcting for guessing, as is done in the 3P model, this procedure eliminates the *noise* guessing creates. Noise is a term used in statistics to contrast with the concept of *signal*, where the former dilutes the latter. As a result, information for every person is used, but only where one can be reasonably sure it is valid information. Waller (1976) applied the ARRG procedure using the Rasch model, and found that item locations and locations of persons who guessed were better recovered with this procedure. Apart from Waller's (1976) limited study and the more recent studies by Andrich, Marais and Humphry (2012, 2016), there seem to be no other studies which investigate the effects of guessing on Rasch model estimates and the effect of procedures like the ARRG on its estimates. In the paper by Andrich et al. (2012) a procedure similar to Waller's is elaborated and applied. It is referred to as a *tailored* analysis. A novel way of testing whether an item has significant guessing is described and applied to both a simulated and an empirical data set in that paper. The procedure is summarized below.

Identifying and Correcting for Guessing

The procedure for identifying and accounting for guessing requires a number of successive analyses. Andrich et al. (2016) described these as *initial*, *tailored*, *origin-equated* and *all-anchored* analyses. The initial analysis is self-explanatory in that it is the first analysis of the set of data in which both the item and person parameters are estimated. The tailored analysis is a form of post-hoc adapting or *tailoring* of a person's proficiency relative to an item's difficulty before administering the item. An item that is considered too difficult for a person is not administered. In the post-hoc tailoring, this involves using the parameters of the initial analysis to eliminate those responses likely to be guessed. Thus from the initial analysis, and based on a person's proficiency estimate and an item's difficulty estimate, if the probability of a response according to the dichotomous Rasch model is less than chance (e.g. 0.25 in the case of 4 alternatives), then this response, whether correct or not, is converted to missing data. Because correctly guessed responses will generate more correct responses than justified for an item based on its difficulty, the item will appear easier in the initial than in the tailored analysis. Because more guessing is likely to occur on more difficult items, the more difficult the item, the greater the increase in its relative difficulty in the tailored analysis compared to the initial analysis.

However, because the sum of the item difficulties is constrained to the same value in a typical analysis, for example 0, the difficulties cannot be compared directly. Thus because the more difficult items will be more difficult in the tailored analysis, and the item difficulty estimates sum to 0, the easier items will be easier. To compare the difficulties from the two analyses, it is necessary to equate the origin of the two analyses. This is carried out in the third analysis in which the mean or average of the difficulties of a few very easy items, which are not expected to be affected by guessing, is fixed to be identical. Because it is considered that the tailored analysis gives the best estimates of the difficulties, this analysis is retained, and initial data is re-analysed with the average of the few easy items equated to their average in the tailored analysis. This is the *origin-equated* analysis. The difficulties of the tailored and the origin-equated analyses can be compared, with the expectation that the greater the difficulty of the item, the greater its difficulty in the tailored analysis.

In the above analyses, and any analysis, it is not known whether a person actually has guessed a correct answer, and for policy reasons, students generally cannot be penalized because they may have guessed an item's correct answer. Therefore, to estimate students' proficiencies in which the item difficulties are not biased by guessing, a fourth analysis is carried out in which the initial data are re-analysed with all item difficulties fixed to those from the tailored analysis. This is the *all-anchored* analysis. The proficiency estimates of the origin-equated and the all-anchored analyses can be compared. As expected, because of guessing, the proficiency estimates of the less proficient students are greater in the all-anchored analysis. However, the proficiency estimates of the more proficient students are also greater in the all-anchored analysis. This arises because the more proficient students receive greater credit for answering correctly the more difficult items, which have a greater difficulty estimate in the all-anchored compared to the origin-equated analysis. The rationale for this effect is described in more detail in Andrich et al. (2016).

These analyses can be carried out routinely in RUMM2030. Following the initial analysis, the tailored analysis can be run in which the user can specify the chance probability value below which a response is converted to a missing response. The origin-equated analysis can be carried out by first saving an anchor file from the tailored analysis with only the easy items saved on it. Then the initial data are re-analysed with the option *Average item anchoring* and the saved anchor file loaded. Now the mean of the easy items will be the same in the new analysis as in the tailored analysis, but all items will have new difficulty estimates. For the all-anchored analysis, all items from the tailored analysis are saved as an anchor file. Then the initial data are re-analysed with the option *Individual item anchoring* and the saved anchor file loaded. Now all item difficulties remain as in the tailored analysis, but each person has a new estimate.

Exercises

Exercise 3: Advanced analysis of dichotomous responses Part B in Appendix C.

References

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*(1), i7–i16.

Andrich, D., Marais, I., & Humphry, S. M. (2012). Using a theorem by Andersen and the dichotomous Rasch model to assess the presence of random guessing in multiple choice items. *Journal of Educational and Behavioral Statistics, 37*(3), 417–442.

Andrich, D., Marais, I., & Humphry, S. M. (2016). Controlling guessing bias in the dichotomous Rasch model applied to a large scale, vertically scaled testing program. *Educational and Psychological Measurement, 76*(3), 412–435.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, Massachusetts: Addison-Wesley.

Waller, M. I. (1973). *Removing the effects of random guessing from latent trait ability estimates.* Unpublished Ph.D. Dissertation, The University of Chicago, Chicago.

Waller, M. I. (1976). *Estimating parameters in the Rasch model: Removing the effects of random guessing (Research Bulletin RB-76-8)*. Princeton, New Jersey: Educational Testing Service.

Waller, M. I. (1989). Modeling guessing behaviour: A comparison of two IRT models. *Applied Psychological Measurement, 13,* 233–242.

Further Reading

Andrich, D., & Marais, I. (2014). Person proficiency estimates in the dichotomous Rasch model when random guessing is removed from difficulty estimates of multiple choice items. *Applied Psychological Measurement, 38*(6), 432–449.