Chapter 26
Analysis of More Than Two Facets and Repeated Measures

In this chapter, we extend the application of the Rasch model from the standard two-facet to a three-facet design. In this extension, the term *facet* is introduced. The standard design of a person-by-item response matrix is said to have two facets, a person and an item facet. The three-facet design has a structure on the items. The three-facet analysis, implemented in RUMM2030 with the Rasch model, parallels this design and can be used for analysing responses involving a judge or repeated measurements. In the chapter, we also describe two other ways that repeated measurement data can be analysed using the standard two-facet Rasch model analysis.

From a Two-Facet to a Three-Facet Rasch Model Analysis

In Chap. 1, we introduced the Rasch model as arising from the requirement of invariant comparisons of persons and items within a frame of reference. In Chap. 19, we gave the example that if two markers assessed student performances, we would require that the assessments are invariant with respect to the marker. In this chapter, we refer to graders, raters, markers and other terms for assessors, with the more evaluative term *judges*. Further, whether the assessment in more than two ordered categories is of the rating or partial credit kind, we refer to it simply as a *rating*.

In many assessment settings, a performance is rated by a judge on several criteria, for example, in the assessment of writing the criteria may include *organization, grammar, spelling* and so on. Research has shown that, even with training, judges can vary in severity of rating (Myford & Wolfe, 2004). Therefore, instead of only two facets, those of person proficiency and item difficulty, a third facet, judge severity, is introduced. Severity can then be quantified on the same scale as the person proficiency and item difficulty, and therefore taken into account.

In many performance assessment designs, a judge is required to assess a performance on multiple criteria. The assessment of essays in educational assessments or health outcomes by a clinician is of this kind. The structure of the design is shown

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_26

Table 26.1 Three-facet design in which *H* judges rate *N* persons on *I* criteria

	Judge			
	1	2	…	*H*
Criterion	1 2.. *i*.. *I*	1 2.. *i*.. *I*		1 2.. *i*.. *I*
Person 1	1 2.. 1.. 0	2 2.. 2.. 1		0 0.. 0.. 0
2	3 1.. 2.. 4	4 2.. 2.. 4		1 1.. 1.. 2
3	0 0.. 1.. 0	1 1.. 1.. 0		0 0.. 0.. 0
.				
.				
n				
.				
.				
N	4 2.. 1.. 3	4 3.. 2.. 4		3 1.. 0.. 2

in Table 26.1 in which each of *H* judges rate *N* persons on *I* criteria. In this case, the combination of a judge and a criterion is effectively an item in the two-facet design. In tables such as Table 26.1, there is very likely to be structurally missing data because it is necessary to have a limited number of judges rate each performance, maybe sometimes only two. This is not a problem if the design includes links in the sense that every performance is assessed by a combination of judges making it impossible to form subsets of performances that are assessed by mutually exclusive sets of judges.

In responses involving assessments in ordered categories, the standard two-facet model structure of Eq. (21.1) from Chap. 21, which characterizes only persons and items, is extended to a three-facet structure (Linacre, 1989; Linacre & Wright, 2002; Lunz, Wright & Linacre, 1990). Specifically, Eq. (21.1) is expanded to the form

$$\Pr\{x_{nih} = x\} = \exp\left[x(\beta_n - \delta_i - \omega_h) - \sum_{k=1}^{x} \tau_k\right] / \gamma_{ni} \qquad (26.1)$$

where ω_h is the severity of judge h, $h = 1, 2, \ldots, H$, and $\Pr\{x_{nih} = x\}$ is the probability that person n obtains a rating x on criterion i from judge h. Again, γ_{nih} is the normalizing factor which is simply the sum of the possible numerators. The parameters β_n, δ_i and τ_k remain the person proficiency, criterion (item) difficulty and threshold difficulty, respectively. In this structure, it is assumed that all judges are consistently more or less severe irrespective of the criterion, and that the categories across the criteria operate in the same way. These assumptions are reflected in the simple additive structure $\beta_n - \delta_i - \omega_h - \sum_{k=1}^{x} \tau_k$. This is the simplest model with a three-facet design and ordered response categories.

The second level of extension of the model's structure is where the judges are consistent across criteria, but when the criteria might have different numbers of categories or the categories operate differently across criteria. Then the thresholds take the subscript i to give τ_{ki} and the structure of the model is $x(\beta_n - \delta_i - \omega_h) -$

$\sum_{k=1}^{x} \tau_{ki}$. The third level of complication is when the judges are not consistent across the criteria with some judges more severe with some criteria than with others. In that case, each judge-by-criterion combination needs to be considered as an item, characterized by a single parameter, say ξ_{ih}, giving $x(\beta_n - \xi_{ih}) - \sum_{k=1}^{x} \tau_{ki}$ as the structure of the model. If the judges are consistent across criteria, but consistently more or less severe relative to each other, then ξ_{ih} specializes to $\xi_{ih} = \delta_i + \omega_h$ and the simpler structure of Eq. (26.1).

RUMM2030 software implements the three facets with the third structure described above, and then specializes it to Eq. (26.1). Thus, it is possible to study whether or not there is an interaction between the judges and either the thresholds, the criteria, or both. Ideally, the additive structure of Eq. (26.1) holds.

In all designs, the person parameter can be conditioned out and the criteria and judge parameters estimated simultaneously but independently of the person parameters. Given the criteria and judge parameters, the person parameters can then be estimated using maximum likelihood or weighted likelihood estimation. These estimates take account of any variation in the severity of judges, which is particularly important where not all judges assess all performances.

Table 26.2 shows item (δ_i) and judge severity locations (ω_h) and SEs, as well as the test of fit details and threshold locations (τ_k), from a three-facet analysis where ten judges rated persons on six criteria (Marais & Andrich, 2011). Because the data were simulated to fit the structure of Eq. (26.1), there is no misfit evident in the fit-residuals, either items or judges. As indicated elsewhere, however, these fit statistics in real data are to be used in conjunction with other statistics that are concerned with model fit.

The model of Eq. (26.1) takes judge severity into account, but several other judge biases have been described (Myford & Wolfe, 2004). One of these is the halo effect, which is the tendency by a judge to assign ratings more similar than justified on different criteria. The halo effect is a violation of local independence. It can usually be detected by the three-facet model and is revealed through judge misfit. However, in Marais and Andrich (2011) a special case of the halo effect is described and it is shown that this halo is not detected by the three-facet model. The paper shows that halo can be diagnosed using the two-facet model, more specifically using a rack or stack design. These analyses are used also to analyse repeated measurement data and are described next.

Repeated Measures

Because total test scores are not necessarily in a constant unit, the measurement of change over time has been a challenge. Using total scores and CTT presents the problem that a small change in an individual's raw score may mean different amounts depending on whether the initial score is extreme or moderate. The same integer change in scores suggests different amounts of change on the variable depending on the location of the pretest score. When the pretest score is very low or very high, then observed score changes are indicative of more change on the variable than when the

Table 26.2 RUMM2030 analysis from a three-facet model in Marais and Andrich (2011)

Criterion	Criterion difficulty δ	SE	FitRes	Judge	Judge severity ω	SE	FitRes	Threshold	Threshold difficulty τ
1	−0.55	0.12	0.35	1	0.46	0.12	−0.30	1	−1.54
2	−0.26	0.11	0.08	2	0.47	0.12	0.56	2	−0.49
3	−0.06	0.11	0.33	3	0.35	0.12	0.48	3	0.51
4	0.08	0.11	0.18	4	0.15	0.12	−0.24	4	1.52
5	0.26	0.12	−0.40	5	−0.04	0.11	0.19		
6	0.53	0.12	0.03	6	−0.04	0.12	−0.11		
				7	−0.14	0.12	0.30		
				8	−0.32	0.11	−0.58		
				9	−0.37	0.11	0.14		
				10	−0.53	0.11	0.42		

pretest score is moderate. Using the Rasch model to convert raw scores non-linearly to measurements helps to overcome this particular concern. There are different ways one can apply the Rasch model to analyse repeated measurements at different time points (e.g. Fischer, 1989; Embretson, 1991; Wright, 1996). These reflect different challenges in measuring change (Marais, 2009).

The three-facet design can be used to analyse repeated measurements by adding the time points as the third facet. If there were two time points, then $\omega_{\text{time 1}}$ would be the mean person location at time 1 and $\omega_{\text{time 2}}$ the mean person location at time 2. The difference would clearly reflect any change.

Data for the two time points can also be analysed in a standard two-facet Rasch model analysis in two ways. First, by treating items used at time 1 and time 2 as distinct (rack design), in which there is a single person parameter across the two times and where the same items may have different parameter values. Second, by treating persons at time 1 and time 2 as distinct with a different parameter value at the two times (stack design). With the rack design, the change is revealed through the item parameter estimates and with the stack design the change is revealed through the person parameter estimates (Wright, 2003). Each design has advantages and disadvantages in diagnosing features of the data, and they should both be used in understanding a data set and in concluding which changes have taken place. For example, the stack design permits studying differential item functioning over time while, as discussed further in the next section, the rack design permits studying response dependence.

Table 26.3 shows graphically the setup of *racked* and *stacked* designs for *N* persons who responded to eight items at two time points. For the rack design, change for the persons is the difference in the mean *item* locations at time 1 and time 2 ($\bar{\delta}_{\text{time 2}} - \bar{\delta}_{\text{time 1}}$). For the stack design, time is included in the analysis as a person factor. Change for the persons is the difference between the mean *person* locations at time 1 and time 2 ($\bar{\beta}_{\text{time 2}} - \bar{\beta}_{\text{time 1}}$).

Repeated Measurements and Response Dependence

Another challenge in measuring change when analysing responses from two time points using the same set of items is that of response dependence. This can be a problem with the data, and is not a problem because of any property of the Rasch model. However, because the model explicitly implies no response dependence, the model can be used to diagnose and control response dependence in assessments across two or more time points. In repeated assessments, response dependence occurs when factors other than the person and item parameters lead to a response to the same item that is more similar at time 2 to time 1 than it would be if only the parameters of the persons and items governed the responses at both times. In this sense, it is analogous to the halo effect. Such dependence can arise because of some idiosyncratic effect of the item which governs the response at both times (for example, a degree of misunderstanding of the item) or where some effect such as memory affects the response the second time. In educational assessment, response dependence is generally con-

Table 26.3 Data design for *N* persons racked and stacked

RACK

Person	Responses	
	Time 1	Time 2
1	24200000	41200000
2	44444201	44333431
3	00000000	01000000
4	44444102	44434224
5	42110000	44100110
6	44433243	44444432
7	43431010	44443113
8	43334000	43342200
9	31100000	21110100
.	.	.
.	.	.
.	.	.
N	44434331	44444320

STACK

Person	Time	Responses
Time 1		
1	1	24200000
2	1	44444201
3	1	00000000
.	.	.
.	.	.
N	1	44434331
Time 2		
1	2	41200000
2	2	44333431
3	2	01000000
.	.	.
.	.	.
N	2	44444320

trolled empirically by having new items created which assess the same variable, and having some common secure items used on the different occasions which act as links between the times of assessment. Controlling for response dependence experimentally can be more challenging with health outcomes assessment where the variables are mostly of the composite form and where it is less easy to create alternative items with specifically defined outcomes.

Being unaware of or failing to control for response dependence can lead to incorrect conclusions, which can be serious when evaluating the effect of a treatment. Response dependence can, depending on initial measurements relative to the person distribution, either reduce or inflate change (Marais, 2009). Olsbjerg and Christensen (2015) and Andrich (2017) have provided a methodological solution for determining change in the presence of response dependence in repeated measurements. The solution is based on the principle developed by Andrich and Kreiner (2010) of quantifying the amount of response dependence between items and requires the data to be racked. This principle was studied in Chap. 14. In a repeated measurement design, the responses to an item at time 2 are resolved into separate items for each response to the same item at time 1.

Exercises

Exercise 7: *Analysis of more than two facets and repeated measurements* in Appendix C.

References

Andrich, D. (2017). Controlling response dependence in the measurement of change using the Rasch model. *Statistical Methods in Medical Research*, 1–17.

Andrich, D., & Kreiner, S. (2010). Quantifying response dependence between two dichotomous items using the Rasch model. *Applied Psychological Measurement, 34*(3), 181–192.

Embretson, S. E. (1991). Implications of a multidimensional latent trait model for measuring change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 184–197). Washington, DC: American Psychological Association.

Fischer, G. H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika, 54*(4), 599–624.

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Chicago: MESA Press.

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement, 3*(4), 486–512.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*(4), 331–345.

Marais, I. (2009). Response dependence and the measurement change. *Journal of Applied Measurement, 10*(1), 17–29.

Marais, I., & Andrich, D. (2011). Diagnosing a common rater halo effect in the polytomous Rasch model. *Journal of Applied Measurement, 12*(3), 194–211.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 518–574). Minnesota: JAM Press.

Olsbjerg, M., & Christensen, K. B. (2015). Modeling local dependence in longitudinal IRT models. *Behavior Research Methods, 47,* 1413–1424.

Wright, B. D. (1996). Time 1 to time 2 comparison. *Rasch Measurement Transaction, 10*(1), 478–479.

Wright, B. D. (2003). Rack and stack: Time 1 vs time 2. *Rasch Measurement Transaction, 17*(1), 905–906.