

# Analysis of a scale hierarchy OR Model Selection by Testing Conjoint Checks

## 1 Background

This work considers the implications of using the methodology described in Domingue (In Press) to further analyze the hierarchy of models discussed in Torres Irribarra and Diakow (2012). The models in Torres Irribarra and Diakow (2012), see Figure 1, support inference at different levels. For example, the MON, IIO, and DM models all allow for the ordering of items and persons. The LCR and RSH models are more powerful in that they allow for differences in quantity. This additional power comes at a cost. Data fit with the latter models must be consistent with axioms known (thanks to Luce & Tukey, 1964) to allow for the creation of an interval scale. In this study, we consider only the axioms of single and double cancellation.<sup>1</sup> The methodology developed in Domingue (In Press) and implemented in Domingue (2012a) describes a procedure for checking the consistency of educational measurement data (dichotomously coded responses to items from persons) with these axioms.

NEED TO EXPLAIN HERE IN GREATER DETAIL HOW THE FIGURE AND THE AXIOMS GO TOGETHER.

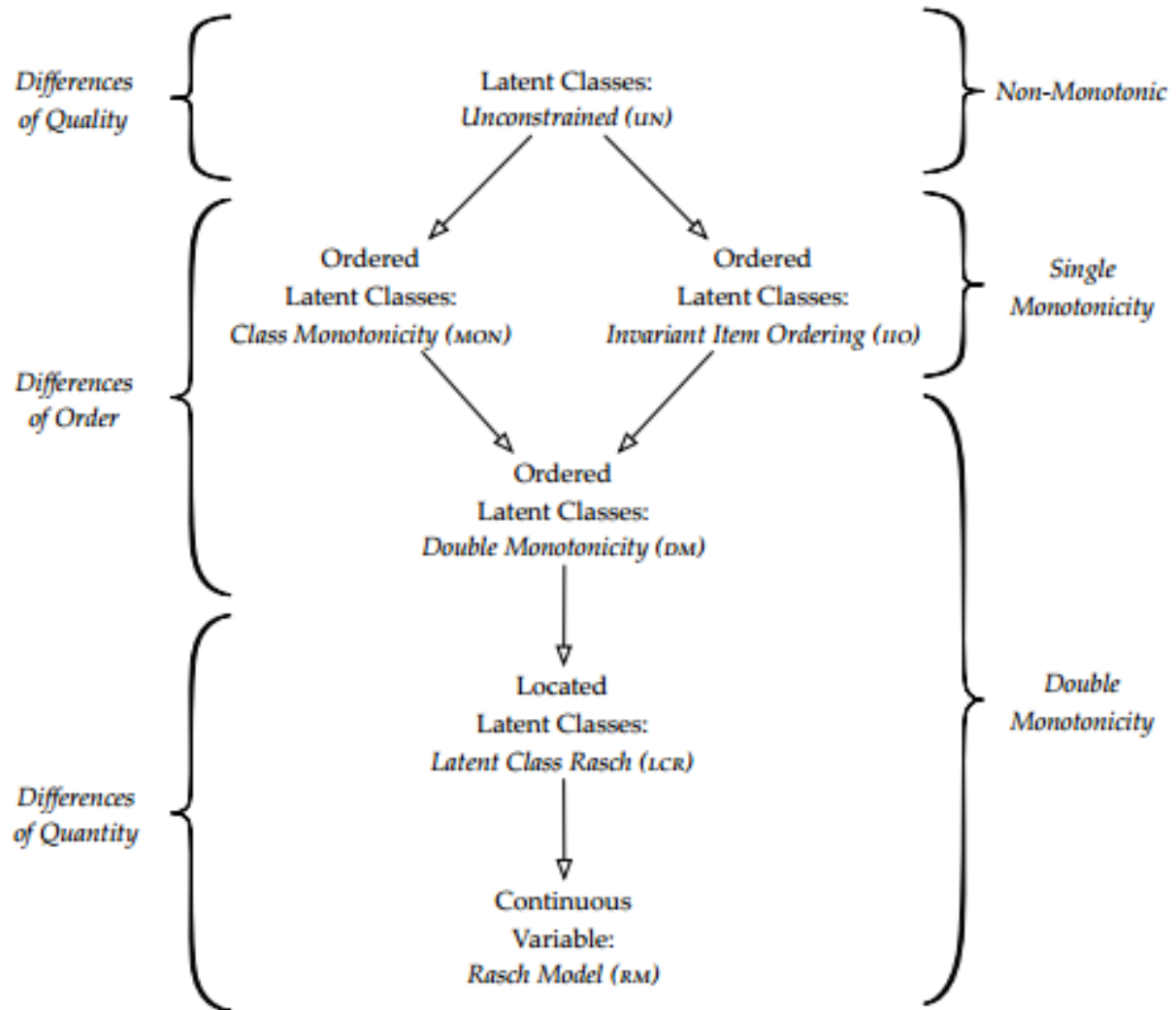
The hierarchy discussed in Torres Irribarra and Diakow (2012) leads to several natural hypotheses regarding violations. These are:

1. The UN model, since it can be fit to purely nominal data, should yield the most violations.
2. At least superficially, there is no reason that there should be substantial differences in the number of detected violations between the MON and IIO models since they each support only a single version of the single cancellation axiom.

---

<sup>1</sup>The higher-order axioms of solvability and the Archimedean condition are typically impossible to verify in practice.

Figure 1: put nice version from original paper here?



25 3. The DM model supports both version of single cancellation but is not necessarily con-  
26 sistent with single cancellation. Hence, it should produce fewer violations than the IIO  
27 and MON models but more than the LCR and RSH models.

28 4. The LCR and RSH models should produce the fewest number of violations given that  
29 they are both consistent with single and double cancellation.

30 Note that from the point-of-view of the hierarchy given in Figure 1, there should be no  
31 difference in the findings with respect to the axioms between the LCR and RSH models.  
32 However, there is clearly an important difference between the two since the former allows for  
33 only a discrete number of classes while the latter allows for a continuum of scores.

34 The key question investigated in this paper is whether these hypotheses hold when the  
35 methodology from Domingue (In Press) is applied to the hierarchy of models. When the  
36 hypotheses do not hold, additional investigations may be in order. For example, it will be  
37 seen that the MON and IIO models produce quite different numbers of axiom violations and  
38 this leads us to ask what it is about these models that causes this difference. Alternatively, a  
39 failure to find evidence that supports the hypotheses related to double cancellation leads us  
40 to argue that checking for double cancellation is probably an intractable problem given the  
41 “messy” nature of the data in question.

## 42 2 Models

43 describe the models in the scale hierarchy.

## 44 3 Results

45 N different data sets were then simulated using each of the six different generating models.  
46 Each of the datasets was then checked. Results are shown in Figure 2. For violations  
47 detected in both randomly formed 3-matrices as well as adjacent 3-matrices, the algorithm  
48 clearly distinguishes between the UNC and MON models on one hand and the IIO, DM, LCR,  
49 and RSH models on the other. The fact that there is a substantial difference between the  
50 IIO and the MON models is surprising given that these models are formally symmetric. It  
51 appears that the checks are sensitive to poorly ordered columns but insensitive to ambiguities  
52 in row ordering, but this is actually too simple an explanation. This issue is explored at length  
53 in Section 3.1.

That the algorithm, which simultaneously checks for both forms of single cancellation as well as double cancellation, can not distinguish between the IIO, DM, LCR, and RSH models is surprising given that IIO and DM do not support double cancellation in general while LCR and RSH do. The checks seem relatively insensitive to violations of double cancellation. Section 3.2 further explores why the checks can not detect violations of double cancellation and also demonstrates that checks for double cancellation are quite challenging given the “messiness” of real-world data and add little to checking simply for both forms of single cancellation.<sup>2</sup>

### 3.1 IIO versus MON

Recall that the IIO model allows for clear ordering of columns while the MON model *should* allow for clear ordering of rows. Figure 3 compares violation proportions for checks testing only row or column versions of single cancellation in adjacently formed 3-matrices. The top panel demonstrates that the checks are relatively insensitive to violations associated with column ordering<sup>3</sup>, but it is reassuring to note that fewer violations are detected with data generated from the IIO model. The bottom panel shows that the MON model generates large numbers of row violations. Note that it is surprising that the IIO model does not since individuals can not be consistently ordered with this model. Given that MON maintains person ordering, in contrast to IIO, the opposite is to be expected. The next several sections are an investigation into why we observe so many more MON violations.

#### 3.1.1 What is the difference between person and item violations?

One potential theory for why row violations are so readily detected is that row “estimates” come with two sources of error:

1. Error in estimates of individual abilities.
2. Error in aggregating rows of common abilities.

---

<sup>2</sup>It may in fact be the case that checking double cancellation in the presence of measurement error is just as challenging as checking the other higher-order axioms!

<sup>3</sup>there is something about this result that i’m not totally comfortable with and it has to do with the fact that with the mon models, the item parameters should be ‘singular’ and yet i’m convinced that they’re being accurately estimated and this leads to the low violations. what if we looked at fit statistics or SEs for the item parameters here? ONE WAY OF TESTING THIS WOULD BE TO VARY THE DIFFERENCES BETWEEN THE DIFFICULTIES OF THE COLUMNS. TIGHTER ITEMS, IN THAT THEIR DIFFICULTIES ARE CLOSER, SHOULD LEAD TO MORE ERRORS (AT LEAST FOR IIO). FOR MON, NEED TO EXPERIMENT WITH HOW BAD THE ITEM VIOLATIONS ARE.

Figure 2: Boxplots showing distribution of violation means (for 3-matrices from randomly chosen and adjacent rows and columns) for the different models

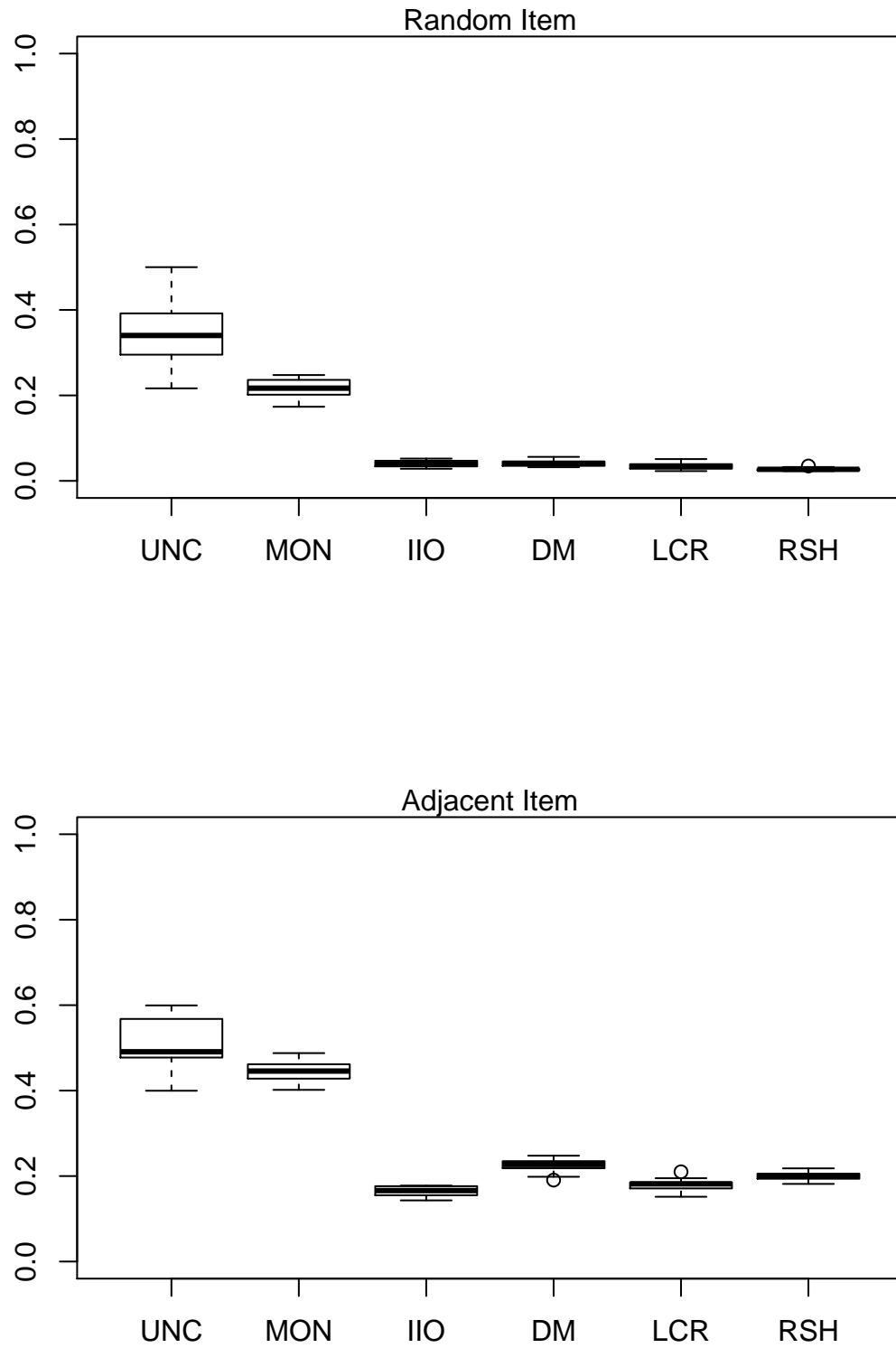
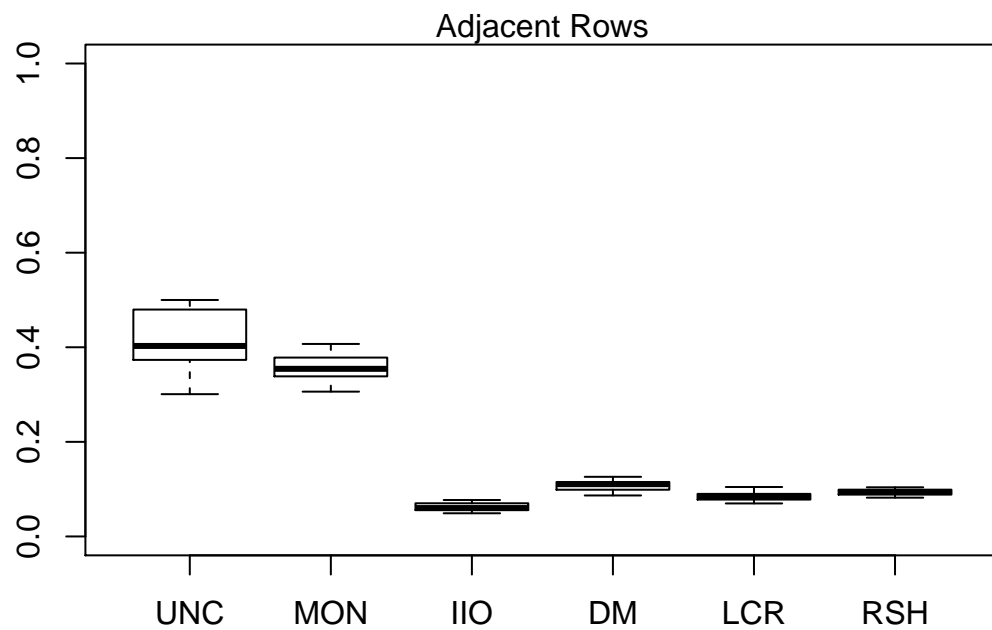
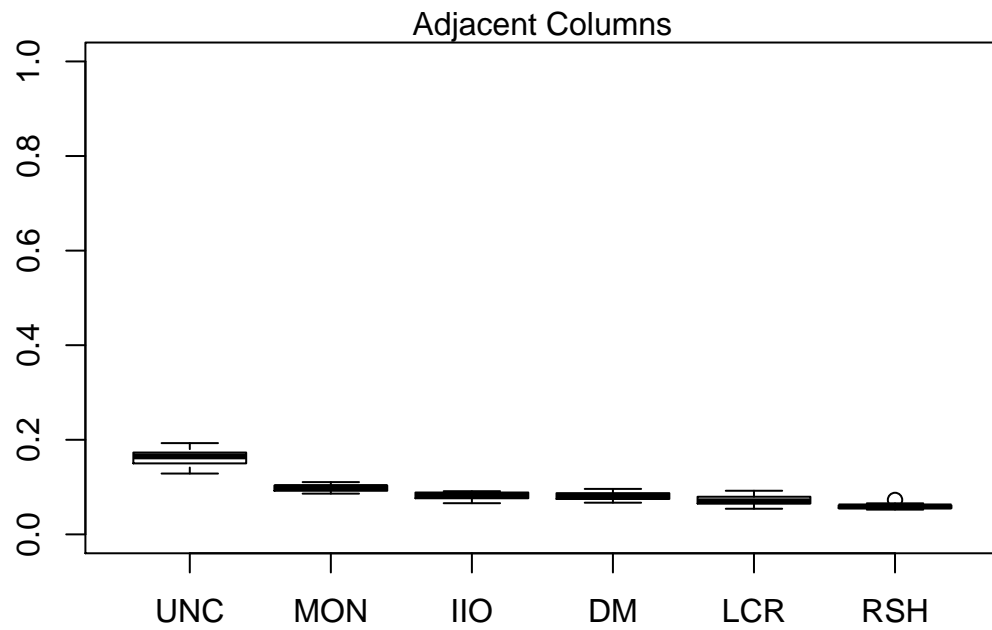


Figure 3: Boxplots showing distribution of violation means for separate checks of single cancellation



In contrast, column orderings only have error due to the error in individual item parameters (which are also more precisely estimated than individual person parameters). Table 1 is a first attempt to make sense of this issue. A simulation involving responses (generated via the Rasch model) from 5000 individuals to a varying number of items was considered. The first form of the test involved only 45 items. With that form, we are generating more violations involving rows (11.6%) than columns (7.3%). When ability estimates are more accurate, estimates are based on responses to 450 items, there is little difference between the violations for the version of single cancellation involving rows and the version involving columns. What is curious is that there has been an uptick in the overall percentage of violations for both forms of single cancellation with this increased precision of ability estimates.<sup>4</sup>

Table 1: Increasing precision of ability estimates

	Rows	Columns
Normal Precision <sup>a</sup>	11.6%	7.3%
Increased Precision <sup>b</sup>	19%	19.2%

<sup>a</sup> 45 items.

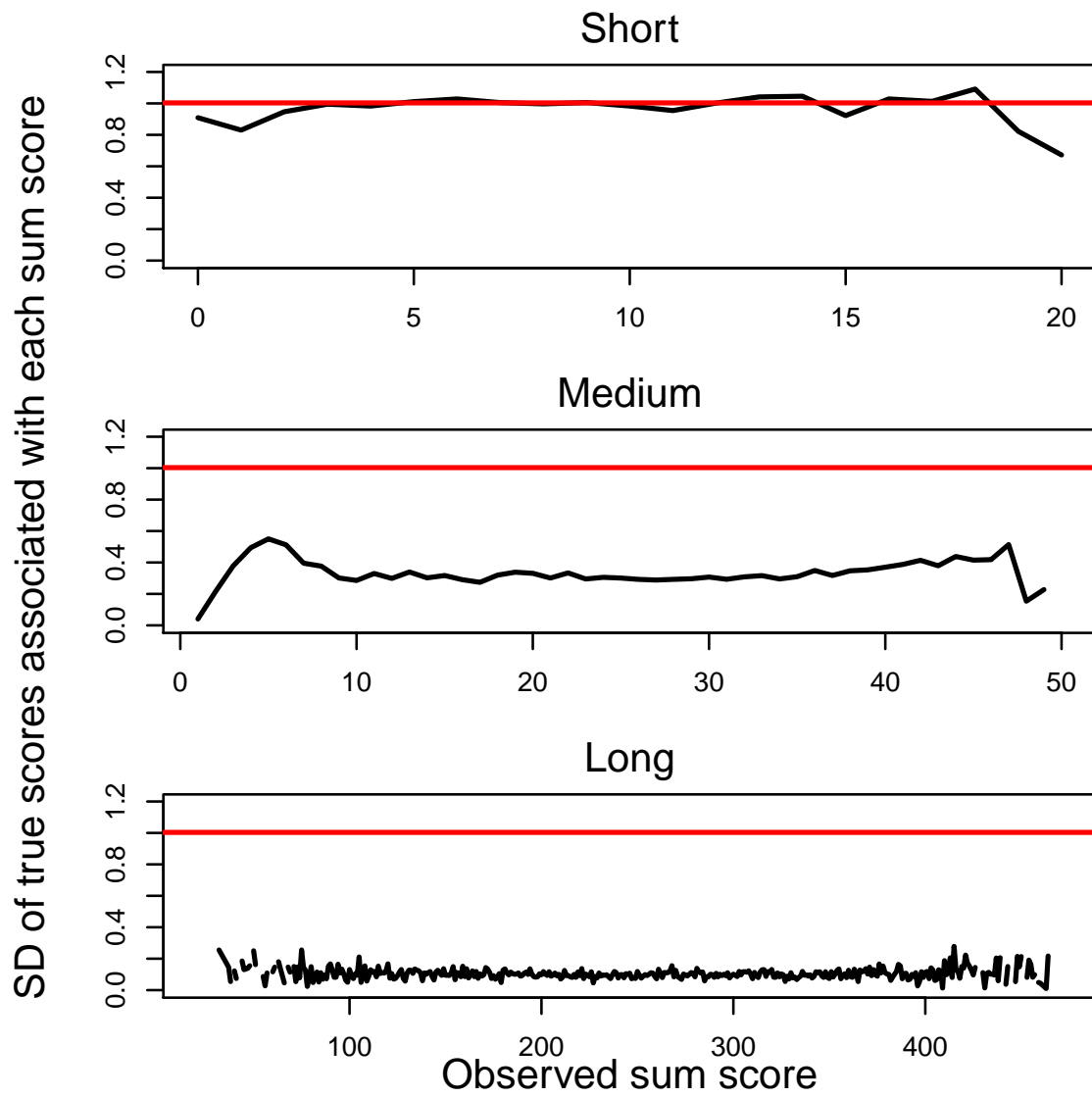
<sup>b</sup> 450 items.

### 3.1.2 Effect of row aggregation on violations?

Consider a series of simulations based on 5000 respondents whose abilities are drawn from a normal distribution and three forms of a test: short, medium, and long. The short form, shown on the top of Figure 4, consists of 20 items. Figure 4 displays the true score SD for all true scores whose simulated response strings yield a given sum score. For the short form, each sum score contains a set of individuals the SD of whose true scores is fairly consistently equal to *or greater than* the SD of all the true scores. This is a different way of looking at measurement error. From this perspective, the sorting done by the test of the respondents into different ability groups is no greater than we would expect than if the respondents were just sorted randomly. The medium length test form, 50 items, has a true score SD for each sum score that is typically around 1/3 of the overall true score SD. The long form, 500 items, is too long to be practical for most applications but is considered as an example. For this form, the true score SD for each sum score is around 20% of the overall true score SD.

<sup>4</sup>One reason for this could be the row aggregation issue discussed below. Another could be the “mesh” issue discussed in my dissertation.

Figure 4: SD in true scores as a function of observed sum scores.





## 3.2 Reconsidering double cancellation

As described in Domingue (In Press), checking for double cancellation is challenging due to the lack of order along the minor diagonals. Evidence presented in the below sections demonstrate that, in fact, there tends to be little difference between the checks with double cancellation, as implemented in Domingue (2012a), and a version that only requires both forms of single cancellation. To illustrate this point, data from Table 2 Perline, Wright, and Wainer (1979) is used.

### 3.2.1 No advantage to double cancellation

Figure 5 compares mean violation rates for the adjacent and random checks when single and double cancellation are checked (denoted double) as opposed to only single cancellation (denoted single). For the random checks, there are essentially no differences. For the adjacent checks, there are small differences for items 3, 5, and 7. For these three, the addition of the double cancellation check adds a small amount of stringency. This is perhaps rather surprising. The next section is an attempt to demonstrate why this is happening.

### 3.2.2 The difficulty of checking for double cancellation

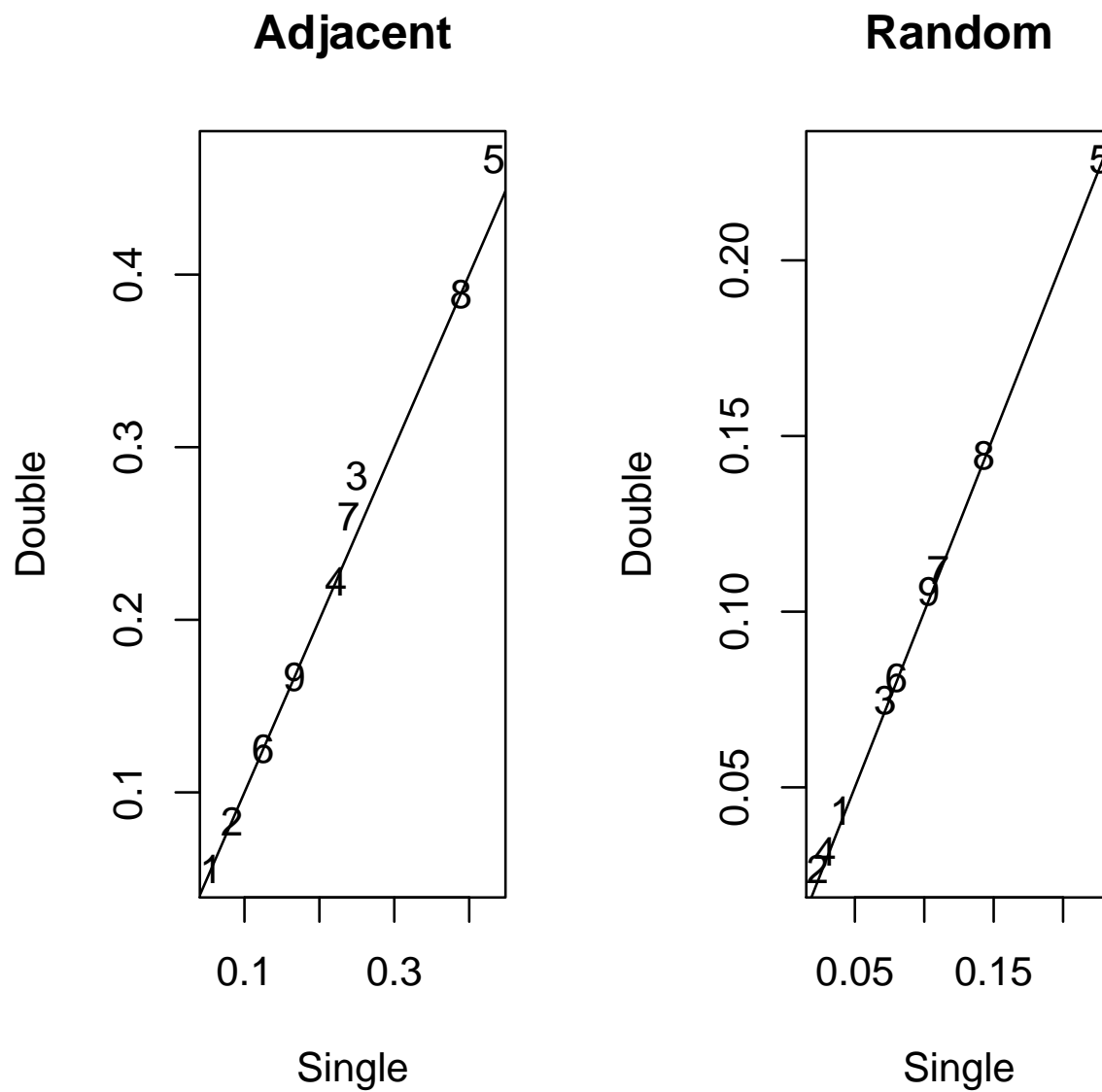
To understand why double cancellation is such a fickle issue with empirical data, it will help to demonstrate via a few empirical examples. Let us begin with rows 3 through 5 of the Perline et al. (1979) data and items 2, 4, and 9. The empirical proportions are:

0.15	0.39	0.67
0.24	0.40	0.70
0.33	0.51	0.78

This 3-matrix was chosen since it is seemingly (in terms of observed data, not true values) in compliance with one form of double cancellation. Focus on the (1, 3) element. Given the fact that  $0.24 < 0.39$  and  $0.51 < 0.70$ , it should be (and is) the case that  $0.33 < 0.67$ . Let us monitor the nature of the checks as iterations of the Markov chain are accumulated. The basic fact is that the addition of double cancellation here is just not especially stringent. As an example, here are the updated values after the first 1000 iterations of the chain:

0.14	0.28	0.62
0.23	0.42	0.65
0.31	0.53	0.86

Figure 5: Comparison of violation means for single and single+double cancellation checks.



In terms of the limits that make up the jumping distribution in the “woven”<sup>5</sup> Metropolis-Hastings algorithm, the upper-right cell needs to be bigger than 0.28 because of single cancellation and, given that the precedent of double cancellation seems to hold, larger than 0.31. This difference is very slight. Checking for only single cancellation is effectively no different.

If we now look at rows 4 through 6 and columns 4 through 6, we observe a much messier scenario:

0.40	0.52	0.51
0.51	0.73	0.68 .
0.58	0.95	0.91

Here we have an example where a consequent of double cancellation regarding the order of the (1, 3) and (3, 1) cells is unnecessary since there is no consistent ordering regarding the pairs (1, 2), (2, 1) and (2, 3), (3, 2). One final example will round out the demonstration. Consider

0.064	0.23	0.51
0.180	0.33	0.61 .
0.524	0.51	0.64

One form of single cancellation seems violated here. However, the violation is so slight that checking for double cancellation on top of single cancellation yields no additional information.

For the data examined here, there are only 5 cases where 3-matrices formed from adjacent cells have empirical proportions of correct responses that correspond to violations of double cancellation and none of these 3-matrices are violations in the sense described in Domingue (In Press). An alternative way of understanding what is going on might be to compare double cancellation within simulated Rasch data. Let  $S_1$  be the truth value of the statement  $\theta_{2,1} > \theta_{1,2}$  and  $S_2$  the truth value of the statement  $\theta_{3,2} > \theta_{2,3}$ . These are the statements that are involved in both forms of the interesting statement of double cancellation. Let  $S_3$  be the truth value of  $\theta_{3,1} > \theta_{1,3}$ . Table 2 tabulates these truth values from a DESCRIBE SIMULATION. Bold numbers in the table indicate violations. Note that there are no violations with the true probabilities from the Rasch model. This is to be expected given that the Rasch model is known to uphold the axioms of additive conjoint measurement. However, with the observed proportions, we observe violations in quite a large number of cases (proportionally) compared to other possibilities. The punchline is that double cancellation is “violated” with observed proportions even when the underlying probabilities are consistent with the axioms of ACM. Why is this? RELATE IT BACK TO SECTION 3.1.2.

Perhaps make point that it is the unevenness of the items and ability levels that leads to

---

<sup>5</sup>I’m introducing this term to deal with the fact that these chains are being updated simultaneously.

Table 2: Double Cancellation with simulated Rasch data

		$S_3$	
	$S_1$	$S_2$	TRUE FALSE
True Probabilities	TRUE	TRUE	5 <b>0</b>
	TRUE	FALSE	9 12
	FALSE	TRUE	8 8
	FALSE	FALSE	<b>0</b> 7
Observed Proportions	TRUE	TRUE	9 <b>9</b>
	TRUE	FALSE	7 5
	FALSE	TRUE	8 3
	FALSE	FALSE	<b>6</b> 2

the disorder on the minor diagonal.

## 4 Discussion

## References

- Domingue, B. (2012a). ConjointChecks: A package to check the cancellation axioms of conjoint measurement. [Computer software manual].
- Domingue, B. (2012b). *Evaluating the equal-interval hypothesis with test score scales*. Unpublished doctoral dissertation, University of Colorado Boulder.
- Domingue, B. (In Press). Evaluating the equal-interval hypothesis with test score scales. *Psychometrika*.
- Luce, R. D., & Tukey, J. W. (1964). Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology*, 1, 1–27.
- Perline, R., Wright, B. D., & Wainer, H. (1979). The Rasch model as additive conjoint measurement. *Applied Psychological Measurement*, 3(2), 237–255.
- Torres Irribarra, D., & Diakow, R. (2012). Impact of instrument quality on the selection of a latent variable model. In *74th annual meeting and training session*. Vancouver, Canada: National Council on Measurement in Education.