Admissible statistics from a latent variable perspective

Zand Scholten, A.

Publication date
2011

Chapter 6

The Interpretation of Interactions Based on Polytomous Items

Misinterpretation is the most deadly of human sins.

Lester del Rey

Abstract

Inferences about interactions can arbitrarily depend on choice of measurement scale. Inferential error can occur when ordinal observed scores, assumed to represent an underlying quantitative variable, are related to the latent values nonlinearly, as is the case in Item Response Theory (IRT) models. Using several IRT models to simulate data, the risk of inferential error for interactions effects based on dichotomous data was assessed to be high when the test was inappropriate targeted and strongly discriminating for large and small sample sizes and several types of interaction effects (Embretson, 1996; Kang & Waller, 2005; Zand Scholten & Borsboom, 2010a). We extended this work for polytomous items using the Graded Response Model (GRM) to simulate data in a small sample, standard ANOVA setting. We also considered the effect of increasing the number of response categories on the risk of inferential error for several types of interaction effects. Results show that inferential error for inappropriately targeted, highly discriminating test is present in data simulated using the GRM. An increase in number of response categories is associated with a decrease in inferential error. Box-Cox transformed observed scores show acceptable levels of inferential error.

6.1 Introduction

The truth value of inferences based on 'inadmissible' statistics, such as parametric tests performed on ordinal data, can depend on the arbitrary choice of numerical scale[1]. A conclusion that reverses with a change in scale, even though both numerical assignments represent the underlying ordinal relations equally well, is meaningless and therefore highly undesirable. This is why basic textbooks teach social scientists that performing a t-test or ANOVA on ordinal data is, strictly speaking, not allowed (Winer, 1971; Shaughnessy et al., 2000; Breakwell et al., 2003). In the same breath however, such textbooks often downplay the actual risk of making an inferential error based on inadmissible statistics, by stating that in practice these tests can be performed without qualm, since ordinal data often provide a sufficiently close approximation to the interval level of measurement. Whether this assurance is warranted however has received relatively little attention.

A conclusion concerning two group means based on ordinal scores can be reversed by a transformation according to a monotonically increasing function (leaving order intact) that stretches or condenses a certain part of the scale. This can bring the means closer together or further apart, or even change their order. Some interesting examples are provided by Townsend and Ashby (1984), Hand (2004) and Zand Scholten and Borsboom (2010a). Such a change in conclusion can only be arrived at under certain conditions however. When the sample distributions in the two groups have the same shape, they are stochastically ordered, which means it is impossible to find a monotonically increasing function that will change the inference based on these scores (Davison & Sharma, 1988). Early simulation research into the effect of several nonlinear transformations on sensitivity to inferential error in the t-test showed the risk was small (Baker et al., 1966). The same unfortunately does not apply to more complicated comparisons of group means however.

A more complicated type of comparison that runs a much higher risk of producing inferential errors is the interaction effect considered in standard factorial designs. Loftus (1978) was the first to point out that when we assume the data represent a latent quantitative variable and the relation to

[1]The next six paragraphs treat the concept of admissible statistics similarly to previous chapters and may therefore be skipped if the reader so wishes.

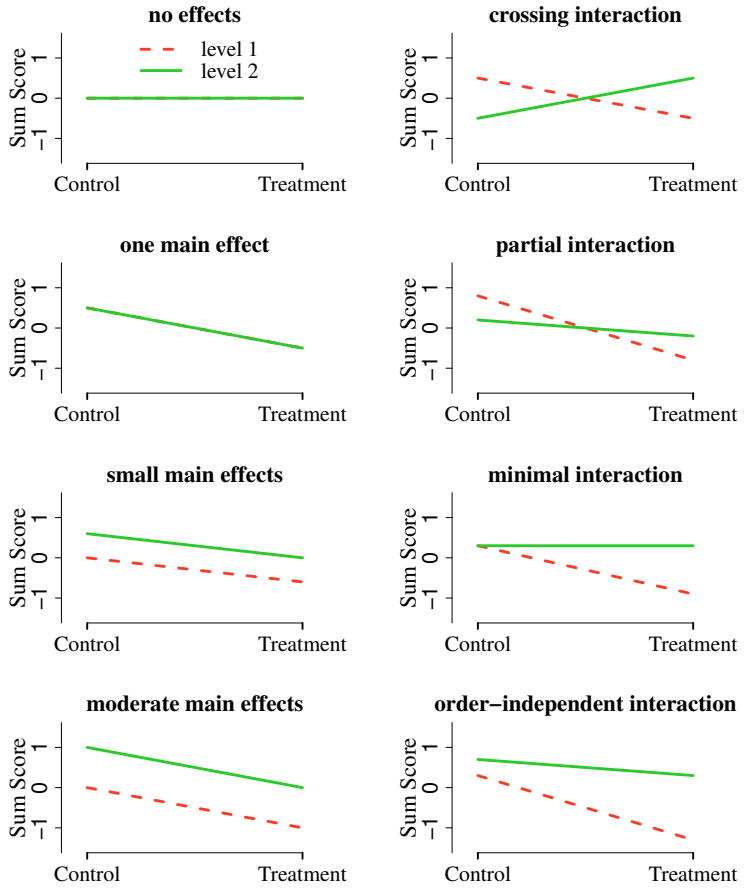
the ordinal observed scores is some unknown nonlinear function, then certain interaction effects are at risk for inferential error while others are not. When the effect consists of no effects, one main effect, a crossing interaction or what we term a partial and minimal interaction (see Figure 6.1), then it is impossible to find a transformation that will change these effects. However, when there are two main effects or an order-independent interaction (see Figure 6.1) it is possible to find a transformation that will change the former into the latter or vice versa, thereby causing an inferential error.

It is important to note that Loftus (1978) did not take measurement error into account, and therefore considered equal means, which cannot be made to differ by any one-to-one transformation, safe from inferential error. In practice, two or more means will never be exactly equal, leaving the risk of inferential error undetermined. In order to narrow down the conditions that present a risk of inferential error, an effort was made to investigate whether standard ANOVA assumptions of normality and homogeneity of variances, resulting in equivalently shaped sample distributions, would preclude inferential error for a two-by-two interaction design (Davison & Sharma, 1990). Unfortunately this attempt was not successful.

Embretson (1996) was the first to employ Item Response Theory (IRT) modeling to investigate a specific type of relation between assumed quantitative latent values and ordinal observed scores. Data were simulated according to the Rasch model, which results in a logistic-type relation between latent and observed scores. Parameters such as item difficulty and discrimination, that influence the linearity of this relation, were manipulated to gauge the risk of inferential error, and to assess the usefulness of IRT ability estimates to mitigate inferential error. Only when the test was inappropriately targeted to the sample, which corresponds to a less linear function, did spurious interactions occur. Latent ability estimates were successfully used to lessen the risk of inferential error. Extension of this line of investigation showed similar results for data simulated using the 2-Parameter Logistic (2PL) IRT model (Kang & Waller, 2005).

Simulation research based on IRT modeling constitutes a great step forward in fleshing out the conditions associated with an elevated risk of inferential error. However, this research (Embretson, 1996; Kang & Waller, 2005) uses a Moderated Multiple Regression (MMR) approach with large sample size, only considers the order-independent interaction and presents IRT abil-

Figure 6.1: Four additive and multiplicative types of effects, categorized according to group order restraints



ity estimation as a solution to the problem of inferential error. Zand Scholten and Borsboom (2010a) approach the problem from the perspective of experimental researchers who often only have access to small sample sizes and fixed effect designs. Estimation of latent abilities using IRT is not an option in this case. IRT modeling can still be useful in this context however, since it allows us to make coherent assumptions about the function that relates latent and observed scores. Using the 2PL model to simulate data, Zand Scholten

and Borsboom (2010a) showed that for small to moderate sample sizes too, the risk of inferential error for inappropriately targeted tests is sizeable. Inferential error occurred for order-independent interactions, but also for other types of effects, namely two moderate main effects, as well as the partial and minimal interaction, albeit to a lesser extent. An alternative solution to IRT estimation of latent abilities was found in a normalizing transformation: The use of normally transformed scores showed a marked decrease in inferential error.

A natural extension of previous research employing IRT modeling is to investigate the risk of inferential error for tests that consist of polytomous items. When more response categories are available allowing persons to provide a more fine-grained response, it is possible that the relation between assumed quantitative latent values and ordinal observed scores will become more linear. To investigate this for the MMR, large sample size case, Morse (submitted) simulated data according to the Graded Response Model (GRM) using five response categories. As in previous research, inferential error is again present for inappropriate tests and can be mitigated using ability estimates. The risk of inferential error is somewhat lower however.

In the current paper we aim to extend this research for the small sample size, standard ANOVA case. As before, we considered several other types of interaction effects besides the order-independent interaction. We also investigated the effect of different numbers of response categories. As the possibility to differentiate one's response increases, the relation between latent and observed is expected to become more linear, resulting in less inferential error. We were interested to see whether an increase in response categories could reduce the risk of inferential error to acceptable levels. We also considered the effect of a normalizing transformation on the risk of inferential error, expecting an error decrease. In the next section a general description of the simulation is provided, followed by a more elaborate treatment in the method section.

6.2 Simulation Study

Our goal was to asses the risk of inferential errors based on Likert-scale data under conditions typical for experimental studies. Therefore a two-by-two fixed effects design with small sample size was chosen. The types of effects

considered are the crossing, partial, minimal and order-independent interaction and corresponding main effects depicted in Figure 5.2. Ability values were generated producing these effects at the latent level. These values were transformed into observed scores using the Graded Response Model, one of the simplest polytomous IRT models available. A test of moderate to short length was simulated. The effect of test appropriateness was investigated by varying item discrimination and response category threshold values, which correspond to item difficulty in the dichotomous case.

The effect of test appropriateness on inferential validity was assessed using data that were simulated repeatedly for each combination of factors. The latent scores and the observed scores were subjected to ANOVAs. A discrepancy between the number of significant interactions obtained from the observed scores and those obtained from the latent scores indicates inferential error. Descriptive statistics and a normalizing transformation of the observed scores were considered, to see if these could signal or mitigate an elevated risk of inferential error. The exact simulation setup is discussed in more detail below.

6.3 Method

6.3.1 Simulation of Latent Scores

We simulated an experimental study with two independent factors F_1 and F_2, each with two levels, and one dependent response variable. The simulation consisted of two steps. In the first step latent ability values were generated according to the model $\theta = t + w_1 \cdot F_1 + w_2 \cdot F_2 + w_{12} \cdot F_1 \cdot F_2$, where t was randomly sampled from a standard normal distribution. Factors F_1 and F_2 indicate assignment to one of two levels (contrast coded -1 or 1) on each factor, and w_1, w_2 and w_{12} indicate the weights used to determine main and interaction effects on the latent level. Sample size was set at 120, resulting in group sizes of 30.

6.3.2 Simulation of Observed Scores

In the second step of the simulation, latent scores were transformed into sum scores. Item responses were generated according to the Graded Response Model, which models item responses to more than two categories using a logistic function with an item discrimination parameter and between-category threshold parameters. The between-category thresholds are introduced for each item i in order to obtain response probabilities for each of X_i categories. This means there are $M_i = X_i - 1$ thresholds for item i. Given the latent ability, the probability of a response falling in or above category x on item i is:

$$p^*_{ijx}(\theta) = \frac{exp(\alpha_i(\theta_j - \beta_{ik}))}{1 + exp(\alpha_i(\theta_j - \beta_{ik}))}$$

This equation produces what is called an operating curve, where θ_j denotes the latent ability of person j, simulated in step one. β_{ik} denotes the ability value associated with a .50 probability of responding above threshold k for item i, where $x = k = 1, 2, \ldots M_i$. α denotes the item discrimination of item i, which is the same for all operating curves [2]. The probability $p_{ijx}(\theta)$ for person j of answering in response category x on item i, is obtained by subtraction:

$$p_{ijx}(\theta) = p^*_{ijx}(\theta) - p^*_{ij(x+1)}(\theta)$$

Where $x = 0, 1, \ldots X_i$, and where $p_{ij0}(\theta) = 1$, and $p_{ij(X_i+1)}(\theta) = 0$ by definition. The resulting response category probabilities were then recoded into scores between 1 and X_i. This was achieved separately for each person on each item, by drawing from a multinomial distribution where the response categories formed the classes and the response category probabilities of the person formed the class probabilities. Finally the item responses were added for each person, resulting in sum scores.

6.3.3 Expected Interaction Effects

Either an additive model or a multiplicative model was used to simulate the latent ability values. The additive model contained no interaction term by

[2] Although item discrimination was the same for all items in a particular simulated test, discrimination was treated as a factor in the simulation study and was set to either a high or low value. We therefore retain the subscript.

setting the weight w_{12} to zero. Any interaction on the latent level under this model will be due to sampling error. The number of significant interactions on the simulated sum scores was therefore expected to be small (low Type I error rate). A significant interaction under this model is spurious and indicates inferential error. Under the multiplicative model, an interaction term was introduced by setting the weight w_{12} to a non-zero value. Latent abilities generated under this model will show significant interactions, except for sampling error. A high number of significant interactions are therefore expected on the simulated sum scores (low Type II error rate, high power). A *non*-significant interaction effect under the multiplicative model indicates inferential error.

6.3.4 Types of Simulated Effects

The effects displayed in Figure 6.1 were simulated by choosing the following combinations of weights w_1, w_2 and w_{12}. The weights $(0.0, 0.0, 0.5)$ produce no main effects (additive model) or a fully crossing interaction (multiplicative model). The weights $(0.0, -0.5, 0.3)$ produce one moderate main effect (additive model), or a partial interaction (multiplicative model). The weights $(0.3, -0.3, 0.3)$ result in two small main effects (additive model), or a minimal interaction (multiplicative model). The weights $(0.5, -0.5, 0.3)$ produce two moderate main effects (additive model), or an order-independent interaction (multiplicative model).

6.3.5 Test Characteristics

The number of items was set to 15, to simulate a moderately short test. Item discrimination was low ($\alpha = 1$) or high ($\alpha = 4$). Item response threshold values were based on item difficulty values used in a previous study (Zand Scholten & Borsboom, 2010a), where differently grouped sets of item difficulties were compared. In the first case the item difficulties matched the distribution of abilities and were spaced equidistantly over a wide range of the ability scale, from -2.5 and 2.5, resulting in an appropriately targeted test. In the second case item difficulties were mismatched to the distribution of abilities and equidistantly over a narrow range, between 0.875 and 1.625, resulting in an inappropriately targeted test. Of course these item difficul-

ties pertain to dichotomous items and need to be transformed into response threshold categories. The number of response categories varied from three to seven categories. Threshold values were chosen around the item difficulty of each item. For example, for three categories the lower threshold was set to (difficulty - 1.00) and the upper threshold was set to (difficulty + 0.00). The threshold values were chosen around the item difficulty as follows for each number of categories:

- three categories: -1.00 and 0.00
- four categories: -1.25, -0.50 and 0.25
- five categories: -1.50, -0.75, 0.00 and 0.50
- six categories: -1.75, -1.25, -0.25, 0.25 and 0.75
- seven categories: -2.00, -1.50, -0.75, 0.00, 0.50 and 1.00

6.3.6 Descriptive Statistics and Tests

The current study consisted of 160 levels: 2 (additive/multiplicative model) $\times$ 4 (choice of weights resulting in main/interaction effects) $\times$ 2 (item difficulty sets used to generate thresholds) $\times$ 2 (item discrimination) $\times$ 5 (response categories). Data were simulated a 1000 times for each combination of levels. For each simulation run, an ANOVA was performed on the latent and 'observed' scores. The significance level was set at 0.05. To see if significant interactions exceeded the expected number, Pearson χ^2-tests were performed. These tests are not to be confused with the standard χ^2-test of independence. Due to the considerable number of tests, a Bonferroni correction was applied, resulting in a significance level of 0.001.

Both the observed and latent scores were subjected to Shapiro-Wilks test of normality and Levene's test of homogeneity of variances. Skewness and kurtosis coefficients were also calculated. Non-normality and heterogeneity of variances were expected in the groups with the lowest means, due to a floor effect. A normalizing transformation could provide a way to mitigate the distorting effect of the non-linear relation between latent and observed scores. Normalization could be considered an approximation to the latent scores without having to resort to modeling. Sum scores were therefore transformed according to a Box-Cox transformation. All descriptive statis-

tics and analyses performed on the original scores were also performed on the Box-Cox transformed scores.

6.4 Results

6.4.1 Main Effects Generated by the Additive Model

Results generated under the additive model are presented in Table 6.1. The number of significant interactions on the observed scores is cross-tabulated for all combinations of threshold sets, item discriminations and number of response categories. With the significance level set to .05, a significant interaction due to chance (i.e., a Type I error) is expected to occur in 50 out of a 1000 iterations. The Pearson χ^2-test statistics[3] and p-values are also provided in Table 6.1. We note that the family-wise error rate or significance level for this large number of tests was set at 0.001. The non-significant χ^2 values attest that the number of significant interactions did not differ from the expected count of 50 for any number of response categories, when no effects, only one main effect, or two small main effects were present. High discrimination and an inappropriately targeted test in terms of difficulty and range did not affect the conclusion concerning the group means for these types of effects.

The situation is different when there are two moderate main effects. When the test is appropriate for the sample, Type I error rate stays in check. However, when discrimination is high and the test is inappropriate, Type I error rate increases slightly. The χ^2-test for two moderate main effects was significant for each number of response categories. When there are six and seven response categories, the Type I error rate increases to values of .10 and .09 respectively. For three to five response categories the rate is a little higher still, ranging between 0.20 and 0.14. However, compared to the dichotomous case with similar sample size (Zand Scholten & Borsboom, 2010a), where significant interactions were found in 44% of the simulation runs, this is a much less pronounced increase.

[3]The χ^2 was calculated using both the number of significant and non-significant interactions (1000 - #(significant interactions)). Due to space limitations, only the number of significant interactions is displayed.

Table 6.1: Number of significant interaction effects (Type I errors) under the additive model with N=120; Expected count is 50

	no effects		1 main effect		2 small effects		2 mod. effects	
X = 3	discrimination		discrimination		discrimination		discrimination	
difficulties	low	high	low	high	low	high	low	high
appropriate	53	53	50	59	58	48	56	45
inappropriate	39	57	34	53	51	71	77	204
$\chi^2(3)$	3.96		7.28		10.74		515.92	
p-value	0.266		0.063		0.013		< 0.0001	
X = 4	low	high	low	high	low	high	low	high
appropriate	59	54	43	56	51	60	60	47
inappropriate	53	53	38	44	60	71	81	188
$\chi^2(3)$	2.42		5.58		13.52		423.45	
p-value	0.490		0.134		0.004		< 0.0001	
X = 5	low	high	low	high	low	high	low	high
appropriate	50	50	41	58	54	59	56	55
inappropriate	52	44	52	58	62	68	80	135
$\chi^2(3)$	0.84		4.48		11.89		172.34	
p-value	0.839		0.214		0.008		< 0.0001	
X = 6	low	high	low	high	low	high	low	high
appropriate	38	52	54	41	44	51	55	47
inappropriate	43	46	48	46	46	61	64	104
$\chi^2(3)$	4.48		2.46		3.66		66.23	
p-value	0.214		0.482		0.300		< 0.0001	
X = 7	low	high	low	high	low	high	low	high
appropriate	38	50	51	55	44	49	52	50
inappropriate	47	48	57	52	42	49	66	86
$\chi^2(3)$	3.31		1.66		2.15		32.76	
p-value	0.347		0.645		0.542		< 0.0001	

6.4.2 Interaction Effects Generated by the Multiplicative Model

Under the multiplicative model, the number of significant interactions corresponds to the power of the test. Power is expected to be relatively high.

Because power is influenced by effect size, and since effect size varies between conditions, a reasonable expected count for each type of effect needed to be determined. Therefore the number of significant interactions on the appropriate, matched wide test, averaged over high and low discrimination, was used as a baseline for each type of effect. The number of significant interactions, expected counts and test results obtained under the multiplicative model[4], are displayed in Table 6.2.

Power is very high and unaffected by an inappropriate test with high discrimination when the effect consists of a crossing interaction, no matter the number of response categories. The same seems to apply to the partial interaction. Although the χ^2-test is significant for each number of response categories for this type effect, the drop in power is not very great and actually goes up when discrimination is high. Overall differences between high and low discrimination are greater than differences between the appropriate and inappropriate test. For the minimal interaction a similar pattern emerges when there are seven response categories. When there are three to six response categories, the power dropped more than 10% if the test is inappropriate and discrimination is high versus low. The decreased power varies between .68 at its worst and .83 at its best, increasing with the number of response categories.

Compared to data generated under the additive model, this is nowhere near as extreme as in the dichotomous case, where power dropped to .40 under otherwise comparable circumstances. For the order-independent interaction the situation is somewhat more serious. The number of significant interactions is markedly lower for the inappropriate test compared to the appropriate test, especially when discrimination is high. Power varies between .38 and .74, once more increasing with the number of response categories. Compared to the dichotomous case however, the decrease is less impressive. With the same simulation generating dichotomous data, power dropped to .09 (!). We also see a change in direction of the effect of discrimination. High

[4]In several cases the expected count for non-significant interactions (not displayed in Table 6.2) was zero. In these cases the expected count was changed to 0.1. Another problem was the occurrence of empty or very low cell counts for the non-significant columns. In these cases Fisher's exact test was performed to obtain p-values. Test statistics are not provided if this test was performed. Fisher's exact test does not compare observed scores to a specific expected count, but takes the expected count from the marginals as a standard χ^2-test of independence.

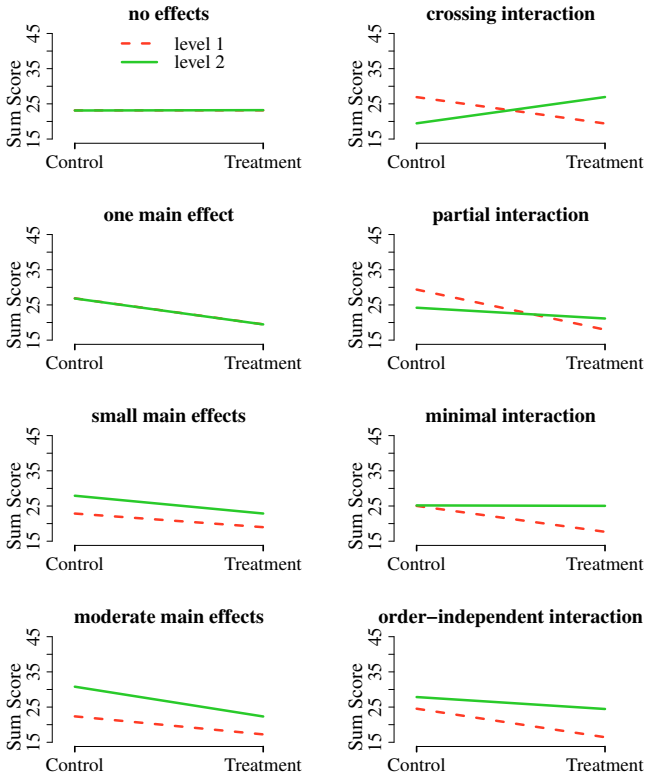
Table 6.2: Number of significant interaction effects (Type II errors/power) under the multiplicative model with N=120

	crossing		partial		minimal		order-indep.	
X = 3	discrimination		discrimination		discrimination		discrimination	
difficulties	low	high	low	high	low	high	low	high
appropriate	998	1000	758	875	770	902	774	902
inappropriate	997	999	779	817	710	682	634	376
exp. count	-		816		836		838	
$\chi^2(3)$, p-value	- , 0.538		54.71, < 0.0001		352.32, < 0.0001		1939.16, < 0.0001	
X = 4	low	high	low	high	low	high	low	high
appropriate	997	1000	777	890	790	892	802	912
inappropriate	1000	1000	796	833	754	730	612	440
exp. count	-		834		841		857	
$\chi^2(3)$, p-value	- , 0.341		56.56, < 0.0001		187.65, < 0.0001		1958.07, < 0.0001	
X = 5	low	high	low	high	low	high	low	high
appropriate	1000	1000	762	884	794	906	780	913
inappropriate	997	999	797	847	752	765	618	522
exp. count	-		823		850		846	
$\chi^2(3)$, p-value	- , 0.736		59.68, < 0.0001		181.18, < 0.0001		1272.64, < 0.0001	
X = 6	low	high	low	high	low	high	low	high
appropriate	998	1000	774	890	801	916	795	911
inappropriate	999	1000	808	872	752	800	653	661
exp. count	-		832		858		853	
$\chi^2(3)$, p-value	- , 0.617		63.70, < 0.0001		174.11, < 0.0001		666.65, < 0.0001	
X = 7	low	high	low	high	low	high	low	high
appropriate	996	1000	788	897	802	904	815	913
inappropriate	997	1000	789	882	771	833	717	741
exp. count	-		842		853		864	
$\chi^2(3)$, p-value	- , 0.072		77.80, < 0.0001		98.30, < 0.0001		353.52, < 0.0001	

discrimination leads to lower power when the number of response categories lies between three and five. For six or seven response categories, discrimination does not seem to affect power in the ill-matched test.

To give some idea of what the simulated observed effects looked like, the observed means were averaged over all iterations and plotted in Figure 6.2. The plotted effects are based on data simulated using three response categories. Considering that the three response category data were the most erratic compared to other numbers of categories, the observed effects in Figure 6.2 resemble the true effects in Figure 6.1 very closely.

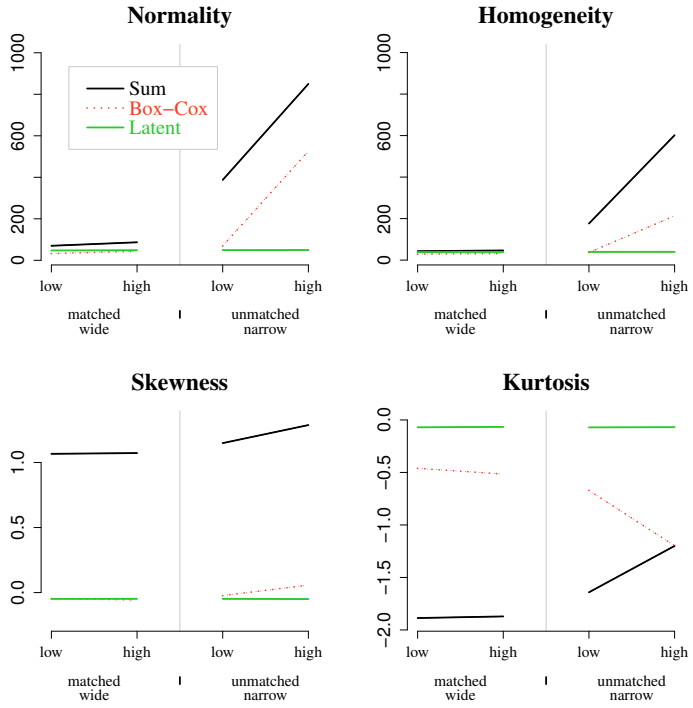
Figure 6.2: Mean observed effects for unmatched narrow test with high discrimination and three response categories. Sample size is 120



6.4.3 Distributional Shape of Observed and Box-Cox Transformed Scores

Deviations from normality and homogeneity of variances could indicate an increased risk of inferential error. Normality and homogeneity of variances assumptions were therefore tested for all combinations of effect types, number of response categories, discrimination and threshold sets. Skewness and kurtosis coefficients were also calculated. All the tests and statistics were calculated for the original scores and the Box-Cox transformed observed scores. Figure 6.3 shows the number of significant Shapiro-Wilks and Levene's tests and the mean skewness and kurtosis, averaged over effect type and number of response categories.

Figure 6.3: Rejection rate for Shapiro-Wilks and Levene's test and standard Skewness and Kurtosis coefficients for sum scores, Box-Cox sum scores and latent scores



The same results, based on the latent ability scores can serve as base rate comparison. In general, violation of normality and homogeneity of variances, and the presence of skewness and kurtosis are associated with inappropriate tests that show high discrimination. The Box-Cox transformed sum scores do much better in this regard.

6.4.4 Box-Cox Transformed Scores and Group Differences in Reliability

A normalizing transformation could be used to approximate the latent scores without having to resort to IRT modeling. An added benefit could be the minimization of inferential error. ANOVAs were therefore performed on the Box-Cox transformed scores. Tables 6.3 and 6.4 show the main results when data were simulated using the additive and multiplicative model respectively.

When compared to the results based on the original scores, it is clear that the Box-Cox transformation is beneficial in all cases. The Box-Cox transformed scores show no inferential error when there are two moderate main effects. There is still some inferential error when there is a minimal or order-independent interaction but the number of inferential errors is roughly cut in half compared to the untransformed sum scores. Pearson's χ^2-tests were no longer significant for any number of response categories on any of the types of additive effects.

Table 6.4 shows that the partial, minimal and order-independent interactions still showed significant deviations from the expected count. This is probably accounted for by the choice of expected number of significant interactions that was not ideal due to a relatively large overall difference in significant interactions between low and high discrimination. Only for three response categories in the minimal interaction case and three and four response categories in the order-independent interaction case, did the power in the Box-Cox transformed scores drop more than 10%.

Table 6.3: Number of significant interaction effects (Type I errors) under the additive model based on Box-Cox transformed scores with N=120; Expected count is 50

	no effects		1 main effect		2 small effects		2 mod. effects	
X = 3	discrimination		discrimination		discrimination		discrimination	
difficulties	low	high	low	high	low	high	low	high
appropriate	52	51	49	61	57	50	48	37
inappropriate	35	55	32	50	54	48	62	64
$\chi^2(3)$	5.37		9.39		1.45		10.8	
p-value	0.147		0.025		0.693		0.013	
X = 4	low	high	low	high	low	high	low	high
appropriate	57	54	43	59	52	61	55	36
inappropriate	56	57	42	38	56	54	41	62
$\chi^2(3)$	3.16		7.12		3.73		9.39	
p-value	0.368		0.068		0.293		0.025	
X = 5	low	high	low	high	low	high	low	high
appropriate	51	51	41	60	49	53	50	44
inappropriate	53	53	51	64	49	51	63	46
$\chi^2(3)$	0.42		7.96		0.25		4.65	
p-value	0.936		0.047		0.969		0.199	
X = 6	low	high	low	high	low	high	low	high
appropriate	39	54	52	39	41	52	54	41
inappropriate	44	50	46	47	49	63	44	42
$\chi^2(3)$	3.64		3.16		5.37		4.15	
p-value	0.303		0.368		0.147		0.246	
X = 7	low	high	low	high	low	high	low	high
appropriate	42	52	49	55	40	46	51	44
inappropriate	50	51	62	55	45	48	56	40
$\chi^2(3)$	1.45		4.11		3.05		3.64	
p-value	0.693		0.250		0.384		0.303	

Table 6.4: Number of significant interaction effects (Type II errors/power) under the multiplicative model based on Box-Cox transformed scores with N=120

	crossing		partial		minimal		order-indep.	
X = 3	discrimination		discrimination		discrimination		discrimination	
difficulties	low	high	low	high	low	high	low	high
appropriate	998	1000	754	877	761	894	741	877
inappropriate	998	1000	771	838	760	781	749	665
exp. count	-		816		836		838	
$\chi^2(3)$, p-value	- , 0.388		67.10, < 0.0001		129.76, < 0.0001		359.32, < 0.0001	
X = 4	low	high	low	high	low	high	low	high
appropriate	1000	1000	777	894	779	879	763	870
inappropriate	998	999	796	852	790	820	701	747
exp. count	-		834		841		857	
$\chi^2(3)$, p-value	- , 1.000		62.24, < 0.0001		62.29, < 0.0001		370.79, < 0.0001	
X = 5	low	high	low	high	low	high	low	high
appropriate	1000	1000	759	886	785	896	752	881
inappropriate	998	999	805	867	796	855	727	820
exp. count	-		823		850		846	
$\chi^2(3)$, p-value	- , 1.000		70.88, < 0.0001		72.8, < 0.0001		191.11, < 0.0001	
X = 6	low	high	low	high	low	high	low	high
appropriate	998	1000	771	892	794	905	760	871
inappropriate	998	1000	808	885	792	872	760	864
exp. count	-		832		858		853	
$\chi^2(3)$, p-value	- , 0.366		76.59, < 0.0001		89.11, < 0.0001		141.5, < 0.0001	
X = 7	low	high	low	high	low	high	low	high
appropriate	996	1000	784	897	791	892	778	881
inappropriate	997	1000	790	890	806	884	804	883
exp. count	-		842		853		864	
$\chi^2(3)$, p-value	- , 0.082		85.67, < 0.0001		68.07, < 0.0001		99.11, < 0.0001	

6.5 Discussion

The results show that although a Likert-scaled, inappropriate test is still associated with the risk of inferential error, this risk is much smaller than when dichotomous items are used. Furthermore, the risk decreases as the number of response categories increases. When an inappropriate test is used that consists of items with seven response categories, the only problematic effect types are the order-independent interaction and the double main effect. A Box-Cox transformation provides a substantial improvement in these cases.

Of course it is important to note the limitations of our simulation setup. We chose a moderately short test and small sample size to simulate typical conditions for an experimental setting. The inappropriateness of the test was fixed by choosing specific values for the response category thresholds. The aim was to facilitate comparison to previous results based on dichotomous responses generated by the 2-parameter logistic IRT model. A drawback is that we did not sample threshold values. The threshold values also resulted in extremely hard items, perhaps harder than we typically see in research settings where inappropriate tests cannot be avoided. This leads us to the optimistic conclusion however, that the risk of inferential error in these applied settings is therefore likely to be even smaller.

The Box-Cox transformation again proved to mitigate inferential error, as it did with dichotomous data. For all types of effects, for both types of test appropriateness and for each number of response categories, the Box-Cox transformation reduced inferential error to acceptable levels. The only exception was when the true effect concerned an order-independent interaction and the test was both inappropriate and consisted of a small number of response categories. A normalizing transformation such as the Box-Cox transformation seems to adequately approximate the latent scores in small experiments where IRT modeling is not an option.

Assuming data are generated according to a process that can accurately be described by the GRM, these data are only at risk for producing inferential errors under very specific circumstances. A sizable risk only occurs when the test is highly inappropriate and the number of response categories is low. This does not mean that appropriately difficult items or a high number of response categories will ensure that risk of inferential error will be absent. An important assumption here is that data behave as if generated by the

GRM. Whether this model provides an accurate description for many tests is questionable. Additional work needs to be done on more complex models that a can be associated with much more non-linearity in the test response curve that ultimately determines the risk of inferential error.