Admissible statistics from a latent variable perspective

Zand Scholten, A.

Publication date
2011

Citation for published version (APA):
Zand Scholten, A. (2011). *Admissible statistics from a latent variable perspective*.

Chapter 5

How to Avoid Misinterpretation of Interaction Effects

The only possible interpretation of any research whatever in the 'social sciences' is: some do, some don't.

Ernest Rutherford

Abstract

Inferences about interactions can be invalid due to arbitrary choice of measurement scale. Such inferences are at risk when ordinal observations are assumed to represent a quantitative underlying property that is related to the observed values via a monotonically increasing function as in Item Response Theory. With a simulation study, the risk of inferential error concerning interaction effects is investigated under conditions that are typical for experimental studies in psychology. A standard ANOVA F-test shows inferential errors if the true effect consists of two main effects, an order-independent, minimal or even a partial interaction in some cases. However, inferential error only occurs when test difficulty is ill-matched to the latent abilities in the sample and test discrimination is high. When true interactions are present, this result is more pronounced when sample size is small. When true interactions are lacking, this result is mitigated by larger sample size. Box-Cox transformed observed scores show much less inferential error. Previous results are combined with current results into a flowchart to provide an easy reference.

5.1 Arbitrary Interactions

Experimental psychologists aim to draw general conclusions about psychological properties based on their particular research findings. For example, a higher mean IQ score for Asians, as compared to Caucasians, might be used to conclude that Asians are more intelligent. However, experimental researchers have to be very cautious when making an inference about an underlying psychological variable (intelligence) based on observed results (IQ scores). There are many well-known methodological threats to the validity of such conclusions. One threat that is rarely addressed, even though it can have a serious effect on the substantive interpretation of research findings, occurs when we employ a statistical test that is inappropriate with respect to the measurement level of the dependent variable.

A statistical test is inappropriate if it makes use of either ordinal, interval or ratio level properties, while the observed data represent the underlying variable on a lower measurement level. Such a test can lead to different conclusions, when performed on transformed scores that still correctly represent the underlying property. Tests of interaction effects, which are very common in experimental studies in psychology, are especially vulnerable to this threat.

We present a comprehensive simulation study, investigating important factors that increase the risk of invalid inference concerning interaction effects. We focus on factors that are typical and therefore highly relevant for experimental research. In doing so, we also provide a general reference for applied and experimental researchers, allowing easy identification of factors associated with higher risk for inferential error. Before we can identify these factors, we will start by discussing what makes an inference invalid in the first place[1].

5.1.1 Legitimate Inference

Consider a factitious study into the effect of alcohol on spatial ability. Suppose we compared the number of errors men and women make on a spatial ability test. In the experimental condition, the participant is asked to drink

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

a glass of lemonade beforehand, supposedly to assess the effect of sugar intake on performance. In the experimental condition alcohol is added to the lemonade, the taste of which is concealed using mint oil. The mean number of errors for men and women in the control and experimental condition are presented in Table 5.1a. These results show an interaction effect: women make more errors than men, and they both make more errors in the alcohol condition, but this effect is stronger for women.

In a strict sense, these scores are only ordinal representations of spatial ability. Therefore nothing prohibits us from transforming the scores according to any monotonically non-decreasing function (Stevens, 1946). Table 5.1b shows the means obtained by taking the square root of the scores. These scores represent the ordering of the participants just as well as the original scores. Unlike the original means, however, the transformed means no longer show an interaction. The detrimental effect of alcohol is now the same for men and women. The inference that we draw about the effect of alcohol on spatial ability, depends on an arbitrary choice of measurement scale. Clearly such ambiguity is undesirable.

If we want to ensure our inferences accurately represent the true effects on the underlying property, we need to select a statistical test that produces consistent results under all permissible transformations of the dependent variable. For ordinal scores this means using only nonparametric statistics. This prescribed restriction of the choice of tests was introduced as the theory of admissible statistics (Stevens, 1946), later reformulated into the concept of meaningfulness (Roberts, 1979; Marcus-Roberts & Roberts, 1987) and legitimate inference (Michell, 1986)[2].

Although the need for legitimate statistics has been the subject of much debate (Lord, 1953; Gaito, 1980; Townsend & Ashby, 1984; Zand Scholten & Borsboom, 2009), many basic data analysis textbooks assert the need to take measurement level into account when choosing a statistical test (Winer, 1971; Shaughnessy, Zechmeister, & Zechmeister, 2000; Breakwell, Hammond, & Fife-Schaw, 2003). At the same time, these textbooks also state that the use of inadmissible tests is often justified, since many variables can be assumed to be 'more than ordinal'. Statistical tests that are formally inadmissible are therefore performed with great regularity.

[2] These last two conceptualizations emphasize that not the test statistic itself, but rather the truth value of the conclusion based on the statistic needs to remain invariant.

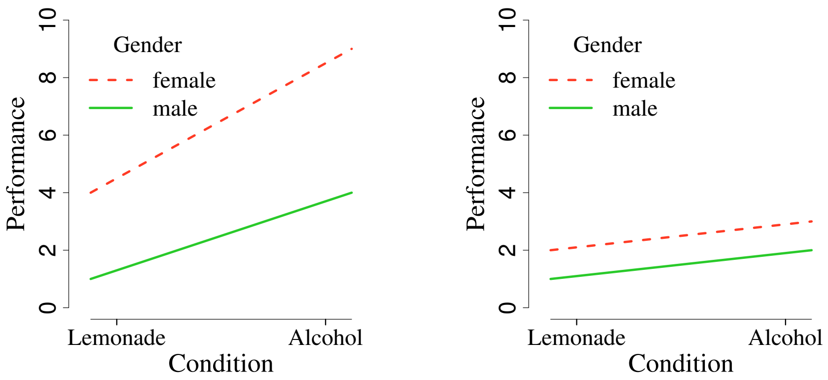
Table 5.1: Mean number of errors for an imaginary alcohol study

(a) Untransformed means

	Condition	
	Lemonade	Alcohol
Men	1	4
Women	4	9

(b) Transformed means

	Condition	
	Lemonade	Alcohol
Men	1	2
Women	2	3



5.1.2 Measurement Level of Observed Scores

The assumption of 'more than ordinal' or interval level measurement is often made for ability and performance tests that consist of 'number correct scores'. Strictly speaking however, such scores are almost always ordinal. We cannot ascertain, for example, whether the difference between an IQ of 100 and 110 constitutes the same increase in intelligence as a difference between 120 and 130. Similarly, it is appealing to conclude that for a cognitive ability test, Jill, who takes 200 milliseconds to solve a simple arithmetic problem, is twice as fast as Jack, who takes 400 milliseconds. Although this statement holds true for the observed ratio-variable 'reaction time', this does not imply that Jill's arithmetic ability, that underlies her observed scores, is also twice as great as Jack's.

To make such statements, one is required to show that arithmetic ability or intelligence has quantitative structure by somehow 'adding' the intelligence of any two people and showing that their combined intelligence can be compared consistently to the combined intelligence of any other two people (Krantz et al., 1971). Of course the concatenation does not have to be direct, or even additive. Few psychological properties lend themselves to direct or indirect forms of concatenation however, which leaves us with at most an ordinal level of measurement in many situations[3]. Without further qualification this means almost all inferences in experimental psychology based on parametric tests, including inferences about interactions based on ANOVA, would have to be considered questionable. Fortunately this conclusion is premature.

5.1.3 Linearity and Inadmissible Tests

For many questionnaires and tests, the advice to treat ordinal scores as interval level scores, and to perform strictly inadmissible tests, is sound. If the relation between the ordinal observed scores and the underlying latent interval variable is linear, the observed scores will produce differences in group means that closely represent the true effects on the latent property (all else being equal)[4]. It is generally assumed that tests of sufficient length, consisting of Likert-scale items with around five to seven response categories will result in a near-to linear relation between latent and observed scores (Dolan, 1994). For a specific subset of tests however, it is not feasible to assume a linear relation between observed scores and the latent property. In IRT it is assumed that for dichotomous items, the relation between the latent variable and the probability of answering an item correctly is best described by a logistic-type function. For the total test score this leads to a nonlinear relation between the observed and latent level.

[3] A noted exception to this rule, at least according to some (Embretson & Reise, 2000; Bond & Fox, 2007), is provided by the Rasch model. This Item Response Theory (IRT) model can be viewed as a form of additive conjoint measurement. This involves the concatenation of items and persons resulting in pairings that correspond to the probability for a person to get the item right. If these pairings can be shown to adhere to a set of axioms (Luce & Tukey, 1964) then interval measurement is said to be achieved. (See Kyngdon, 2008a, Borsboom & Zand Scholten, 2008 and Michell, 2008a, for a discussion concerning this claim.)

[4] It is important to note that the assumption of a linear relation between latent and observed scores presupposes that the underlying property has quantitative structure. This assumption is almost never plausibly supported, empirically or otherwise. See Michell for a detailed discussion on this topic (Michell, 1997).

Figure 5.1: Inferential errors due to nonlinear transformation concerning unequally shaped sample distributions (left) and an interaction (right). Annotated using the alcohol example for illustrative purposes (L = Lemonade condition, A = Alcohol condition)

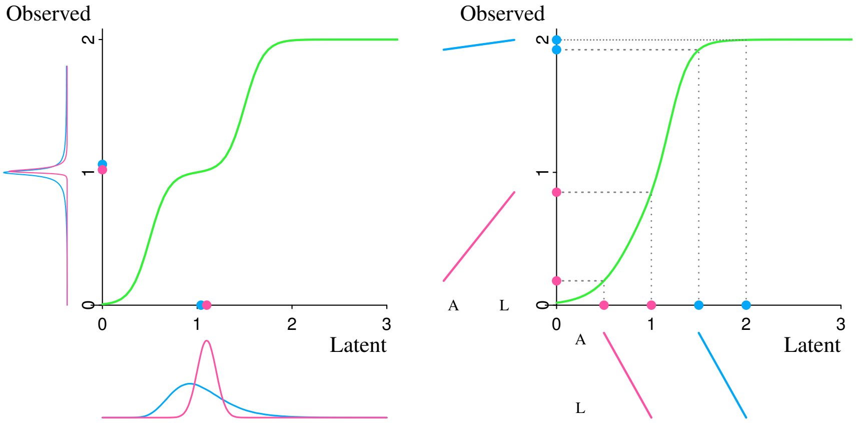


Figure 5.1a illustrates how such a nonlinear relation can lead to an inferential error. The right tail of the latent distribution with the lower mean lies above the right tail of the latent distribution with the higher latent mean. It is possible to find an order-preserving function that will award so much weight to that part of the scale where the first tail exceeds the second, that the first mean is raised disproportionally, and will exceed the transformed second mean. Based on observed scores, we might, for example, infer that Caucasians are more intelligent than Asians. Yet, with direct access to the latent level, we would have drawn the opposite conclusion.

5.1.4 Interpretability of Interactions under Non-Linearity

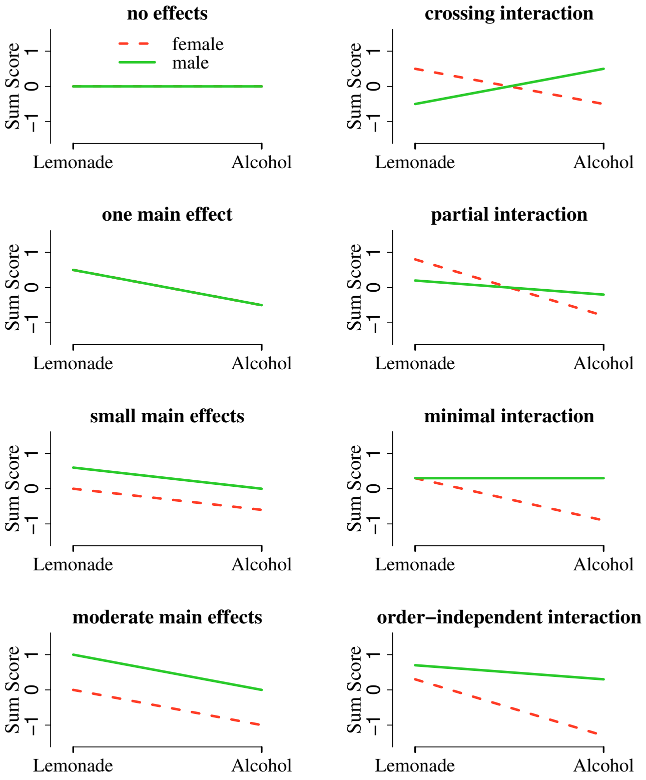
We now come to the interaction effect, which from this point, will be the focus of our attention. For an inferential error to occur that involves reversal of group means, the latent ability distributions of the groups have to be shaped differently. The interaction effect does not require this to result in illegitimate inference, and is therefore even more vulnerable to inferential errors. Figure 5.1b illustrates how an interaction effect can be susceptible to inferential error. Two main effects on the latent variable (parallel lines) are transformed by a logistic function that stretches the lower end of the scale, thereby disproportionally lowering the bottom mean and creating an interaction (non-parallel lines).

If it is plausible to assume that the observed scores are linearly related to the underlying variable, we have little reason for concern. But what if we assume a nonlinear relation? Does this mean all interactions should then be considered uninterpretable? Fortunately, such a drastic conclusion is not necessary. Interactions come in many shapes and sizes, not all of which are equally vulnerable to invalid inference. Assuming all four group means differ, the means can be ordered in 24 different ways. These 24 permutations can be categorized into three classes of effects: crossing interactions, partial interactions and order-independent (main or interaction) effects. When certain group means are equal, additional types of main effects and interactions occur (see Appendix C).

Figure 5.2 displays these effects, annotated in line with the alcohol example to facilitate interpretation. The crossing interaction entails a reversal of the group means between levels of one factor on the other factor (i.e., when sex is plotted on the x-axis the lines still cross). The partial interaction is so named because the group means only reverse for one factor (i.e., when sex is plotted on the x-axis the lines will diverge up and down). The order independent interaction and double main effect are both characterized by the irrelevance of the ordering of the non-extreme means (men in the alcohol condition and women in the lemonade condition). These means can be reversed without affecting the substantial interpretation of the effect. The minimal interaction is named for the minimal requirement needed to create it, namely the deviation of only one group mean.

If nothing is known about the form of the functional relation between latent and observed scores, except that it is nonlinear and nondecreasing, then crossing, partial and minimal interactions will always remain consistent (Loftus, 1978). For these types of interactions, the effect is entirely determined by the ordering of the group means. Uneven stretching of the scale will affect the size of the group differences, but not the substantive interpretation of the effect. The absence of an interaction in combination with exactly one main effect is also unambiguous, as is the total absence of any effect (i.e.,

Figure 5.2: Four additive and multiplicative types of effects, categorized according to group order restraints. Annotated using the alcohol example for illustrative purposes



when all group means are equal). However, effects consisting either of an order-independent interaction, or two main effects with no interaction, are inherently at risk for invalid inference. These types of effects can be transformed into each other by nonlinear stretching or condensing of the scale.

It must be noted however, that Loftus (1978) considered the effect of nonlinear transformation on the exact value of the latent mean. He thereby ignored distortion due to unequally shaped sample distributions, a type of inferential error that was illustrated in Figure 5.1a. The effect of sampling error was also disregarded. Due to sampling error, we may find latent group means that differ from the latent means in the population. This can become a problem when hypothesized effects consist of two or more equal population means. Two values that are equal will remain equal after any one-to-one transformation. However, small differences between group means due to sampling error can be exacerbated by nonlinear transformation, thereby resulting in large differences in observed group means. To our knowledge this problem has not been addressed in the literature so far.

In contrast, the detrimental effect of unequally shaped sample distributions has received much more attention. In the simple two-group comparison or one-way ANOVA case, inference is disambiguated if the groups are normally distributed with homogeneous variances (Davison & Sharma, 1988). Such a restriction preempts the type of invalid inference in Figure 5.1a. However, this approach does not work in factorial ANOVA designs with two or more factors (Davison & Sharma, 1990). Combining the results from Loftus (1978) and Davison and Sharma (1990), we see that the assumptions of normality and homogeneity unfortunately fail to provide a way to exclude inferential error concerning order-independent main or interaction effects. Also, the extent of the detrimental effect of sampling error on effects that incorporate equal means, remains unclear.

5.1.5 Interactions under Specific Non-Linear Assumptions

Besides type of interaction and distributional shape, another factor to consider is the exact form of the relation between the latent and observed variables. When this relation is specified, it is possible to investigate the effect of other research designs and test characteristics on the validity of inferences. Embretson (1996) and Kang and Waller (2005) simulated latent and observed scores using the Rasch model and a two-parameter IRT model, respectively. The impact of main and interaction effect size, item difficulty, item discrimination and test length on the detection of interaction effects was investigated. Inferential errors were found only when the test was too easy or difficult for the simulated sample of persons. This result was more pronounced when test length and effect sizes were greater and the test consisted of highly discriminating items, i.e., items that differentiate well between values of the latent ability.

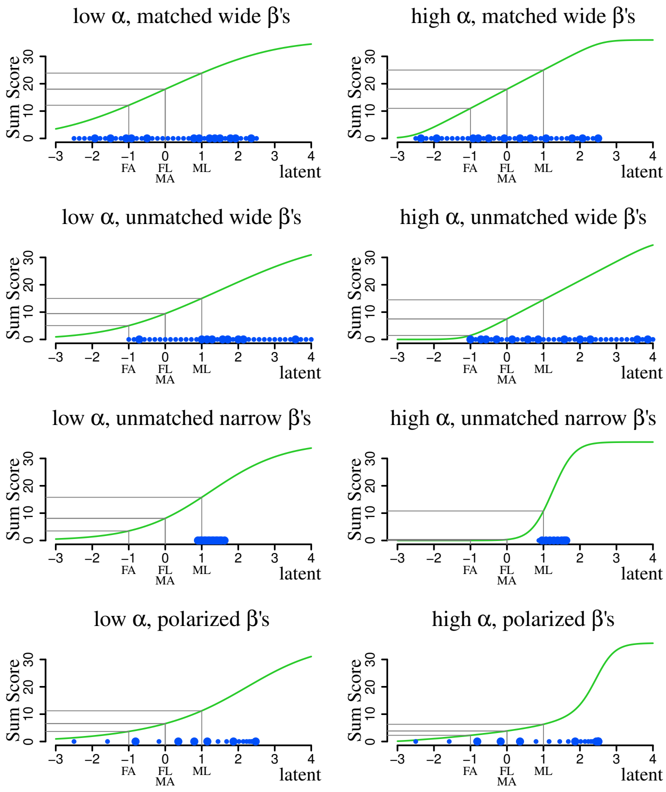
These results make sense when we consider the functional form of the relation between scores on, for example, an inappropriately difficult test and the latent variable. Over the range of the group means, the function will increase slowly for the lowest means, but faster for the higher means (see Figure 5.3 for examples). The means are located where the function is 'least linear' and therefore unequal stretching or condensing of the means is likely, with a higher risk of distorting effects. The longer a test, the more consistent the item responses will be, resulting in an even stronger floor or ceiling effect. The higher the item discrimination, the steeper the slope of the curve, resulting in more sudden and extreme differential stretching and condensing of the scale. The IRT perspective thus allows us to identify when and why the relationship between the latent and observed level is more nonlinear and therefore at risk of producing invalid inferences.

In addition to identifying parameters associated with inferential error, IRT can also help us to avoid inferential error altogether, at least according to Embretson (1996) and Kang and Waller (2005). They advocate the use of IRT modeling to obtain latent ability estimates. These estimates can be analyzed in place of the observed sum scores. Although such advice seems sound in their setting (moderated multiple regression with large sample size), this avenue is often closed to experimental researchers. In many cases the sample size in experimental studies is too small to fit an IRT model. Even if estimates can be obtained, the model still has to show adequate fit. Obviously, latent ability estimates are not going to provide a one-stop, fix-all solution to our inferential problems. More insight is needed into the risk of inferential error under circumstances that are typical for experimental studies. Also, a method to mitigate inferential error when these risk factors cannot be avoided would form a welcome addition to our statistical toolkit.

5.1.6 Identifying and Avoiding Inferential Errors

By considering a specific non-linear relation, we have further reduced the set of inferentially ambiguous interactions: order-independent effects, and possibly minimal interactions (due to sampling error) are at risk for inferential error, but only when the test is mismatched to the abilities of the persons in the sample. Of course this reduction comes at a cost. We have introduced several assumptions in the process, namely that a latent variable exists, that it has quantitative structure and that this structure is represented by the ob-

Figure 5.3: Item difficulties and test response curve associated with the four item difficulty distributions for low and high discrimination. Item difficulties (β's) are displayed as dots on the x-axis. Group means for the additive order-independent effect are included for illustrative purposes (F = Female, M = Male, L = Lemonade, A = Alcohol)



served scores via a fairly specific logistic-type function. Given these assumptions, is there a way to avoid inferential error altogether? The simplest way is to limit our use of measurement instruments to tests of appropriate difficulty and length, consisting of moderately discriminating items. Unfortunately this approach is not always feasible. Even if psychometrically sound tests are available, in many cases it is impossible to use these tests due to the nature of the research subject.

For example, tests used to diagnose psychopathological disorders like depression or schizophrenia need to discriminate well at the high end of the scale to identify people at risk for suicide or sudden aggressive behavior. Such tests will be inappropriately 'difficult' for most of the sample. Another example of a test that is inappropriate by design, occurs when a teaching method is so effective that it renders the test too easy for the participants in the experimental condition while remaining appropriate for those in the control condition. The use of inappropriately difficult tests can also be unavoidable for more substantive reasons, as in the context of stereotype threat, for example. Stereotype threat occurs when the performance of a stereotyped group is lowered compared to the performance of a non-stereotyped group, due to activation of the stereotype (Steele & Aronson, 1995). The detrimental effect of the threatening stereotype on performance of the stereotypes group only occurs when the possibility of confirming the stereotype is real. For this possibility to be perceived, the test has to be sufficiently difficult (O'Brien & Crandall, 2003).

Given that mismatched tests cannot always be avoided, and IRT modeling is rarely a viable option in experimental settings, alternative methods are needed to signal and possibly mitigate an elevated risk of inferential error. With the focus on latent ability estimates, the previous IRT studies (Embretson, 1996; Kang & Waller, 2005) put less emphasis on results pertaining to the distributional properties of the observed scores. Inferential error was found when the test was mismatched and highly discriminating, which entails that the lowest scoring groups on a difficult test (highest for an easy test) will show a floor (ceiling) effect. The scores in these extreme groups will be more heavily skewed and will show less variability. Group differences in descriptive statistics could signal high-risk situations. With standard tests of normality and homogeneity of variances already in place, the required additional effort is minimal.

We therefore conducted a simulation of a typical two-by-two factorial design study. The first aim was to confirm that mismatched item difficulty and high item discrimination would result in inferential errors in a typical small-sample, fixed-effects study. The second aim was to assess the impact of sampling error. Therefore, all effects presented in Figure 5.2, including the minimal interaction, were considered. Third, inferential error was assessed for different sample sizes. Since sample size relates directly to statistical power and generalizability, an effect on inferential error is to be expected. Apart from identifying general factors associated with inferential error, a normalizing transformation was performed in the hope of suppressing inferential error.

5.1.7 Simulation Study

The aim was to investigate the risk of inferential errors for different types of interaction effects under conditions typical for experimental studies. A two-by-two fixed effects design was chosen. The effects generated on the latent variable are the crossing, partial, minimal and order-independent interaction and corresponding main effects depicted in Figure 5.2. Latent ability values were transformed into observed scores using the two-parameter IRT model. A test of moderate length was simulated. Sample size, item difficulty and discrimination were varied systematically.

To investigate whether increasing sample size would decrease the risk of inferential error, we compared a typically small sample size with a larger sample size. Test difficulty was manipulated to produce items that varied in how well they matched latent abilities and to what extent they covered the range of abilities. Discrimination was manipulated to be high or low. To see how changes in sample size, difficulty and discrimination affect inferential validity, data were simulated repeatedly for each combination of factors. ANOVAs were performed on the latent scores and the observed scores. The number of significant interactions on the latent and observed level were compared. Also, several descriptive statistics and a normalizing transformation of the observed scores were considered. Deviations from normality signaled by descriptive statistics and tests could indicate that the risk of inferential error is elevated. A normalizing transformation could help to counter the distortion caused by the nonlinear transformation from latent to observed scores. The exact simulation setup is discussed in more detail below.

5.2 Method

5.2.1 Generating latent and observed 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. First, 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 or 360, resulting in group sizes of 30 or 90 respectively. The first choice reflects sample sizes that are generally found in experimental research.

Second, latent scores were transformed into sum scores. Item responses were generated according to the two parameter logistic model:

$$p_{ij}(+) = \frac{exp(\alpha(\theta_j - \beta_i))}{1 + exp(\alpha(\theta_j - \beta_i))}$$

Where $p_{ij}(+)$ indicates the probability of answering item i correctly for person j. θ_j denotes the latent ability of person j, simulated in step one. β_i denotes the item difficulty and α the item discrimination of item i. 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. The probability $p_{ij}(+)$ was then recoded into a binary score of 0 or 1, by drawing from a Bernoulli distribution with probability $p_{ij}(+)$. Finally the item responses were added for each person, resulting in sum scores.

5.2.2 Spurious and overlooked interactions

Latent abilities were simulated under either an additive model or a multiplicative model. In the additive model the interaction term was taken out of the model by setting the weight w_{12} to zero. Except for sampling error, latent abilities will display no interaction under this model. We therefore expected to find very few significant interactions on the corresponding simulated sum scores (low Type I error rate). Any significant interaction will therefore be designated a *spurious* interaction. Under the multiplicative model, latent abilities were generated with a non-zero weight w_{12}. Barring sampling error, latent abilities generated under this model will show interactions and therefore we expected a high number of significant interactions on the simulated sum scores (low Type II error rate, high power). A *non*-significant interaction effect found under the multiplicative model will be termed an *overlooked* interaction.

5.2.3 Types of simulated latent effects

The effects displayed in Figure 5.2 were simulated by choosing different combinations of weights w_1, w_2 and w_{12}. The weights $(0.0, 0.0, 0.5)$ results in no main effects under the additive model and a fully crossing interaction under the multiplicative model. The weights $(0.0, -0.5, 0.3)$ result in one moderate main effect, and a partial interaction under the additive and multiplicative model, respectively. The weights $(0.3, -0.3, 0.3)$ result in two small main effects under the additive model, and in a minimal interaction under the multiplicative model. Finally we included the weights $(0.5, -0.5, 0.3)$, that result in two moderate main effects under the additive model, and an order-independent interaction under the multiplicative model.

5.2.4 Test characteristics

The test consisted of 36 items to simulate a test of moderate length, thereby approximating the short to moderate test length typically used in experimental studies. Item response probabilities were generated using four differently grouped sets of item difficulties. In the first three cases, the item difficulties were all equally spaced, either between -2.5 and 2.5 ('matched wide'), between -1 and 4 ('mismatched wide'), or between 0.875 and 1.625 ('mismatched narrow'). Finally, the spacing between the item difficulties was varied by polarizing at the extreme of 2.5 and moving in increasing steps, using a power function, towards the extreme of -2.5 ('polarized'). This item distribution was included to simulate tests that are heavily targeted at extreme values at one end of scale, but that do include a few items that cover the rest of the scale to allow for some response variation.

Item discrimination was either low ($\alpha = 1$) or high ($\alpha = 4$). Figure 5.3 displays the test response curve that relates the latent ability to the sum scores and the item difficulty values for the four different sets of difficulties and low and high discrimination. Higher item discrimination corresponds to a steeper test response curve and concentration of item difficulties result in more changes in slope. These changes in slope will be smoothed by low discrimination, resulting in a more linear curve.

5.2.5 Descriptive statistics and tests

Altogether this study consisted of 128 levels: 2 (additive / multiplicative model) × 4 (choice of weights resulting in main / interaction effects) × 4 (item difficulty sets) × 2 (item discrimination) × 2 (sample size). For each combination of levels the simulation was repeated a 1000 times. For each of these simulation runs, an ANOVA was performed on the latent and 'observed' scores, employing a significance level of 0.05. Pearson χ^2-tests were performed to see if the number of spurious and overlooked interactions exceeded the number of significant results expected due to chance. These tests are not to be confused with the standard χ^2-test of independence. Using the Bonferroni correction, a significance level of 0.001 was adopted, due to the large number of iterations and the considerable number of statistical tests.

Shapiro-Wilks' test of normality and Levene's test of homogeneity of variances were performed on the observed and latent scores. Skewness and kurtosis coefficients were also considered. We expected non-normality and heterogeneity of variances, especially in the groups with the lowest means. For these groups the latent scores will be transformed according to the least linear part of the function that relates the latent and observed level resulting in unequal stretching of the scale and a change in sample distribution. In such cases a normalizing transformation could provide a good way to approximate the latent scores without modeling. A byproduct of an attempt to restore the sample distribution to its (assumed) original shape is that unequally stretched parts of the scale are in effect condensed again. Sum scores were therefore transformed according to a Box-Cox transformation. All descriptive statistics and analyses that were performed using the original scores were also performed using the Box-Cox transformed scores.

5.3 Results

5.3.1 Main Effects Generated by the Additive Model

The primary results of the simulation, generated under the additive model, are presented in Table 5.2. Results generated using small and large sample sizes (N=120, 360) are presented in the same table. The number of significant interactions, obtained from a standard ANOVA performed on the simulated observed scores, is cross-tabulated for all combinations of item difficulties and item discriminations. Under the additive model, a significant interaction (i.e., a Type I error) is expected in 50 out of a 1000 iterations. The Pearson χ^2-test statistics[5] and p-values are also provided in Table 5.2. As witnessed by the non-significant χ^2 tests, the number of significant interactions did not differ from the expected count of 50 when no main effects or only one main effect were present. The most extreme deviation constitutes an increase in Type I error of less than 1%. Even when discrimination is high and when the item difficulties are ill-matched to the sample, the conclusion concerning the group means is unaffected by the nonlinear relation with the latent ability.

The situation is different when there are two main effects. Only when the test matches the sample and item difficulties cover the entire ability range, does Type I error rate stay in check. For all three inappropriately targeted tests, increased discrimination leads to more Type I errors. As the test becomes more inappropriate, Type I error rate goes up, in some cases quite drastically. For the least appropriate test (unmatched narrow), Type I error rate increases to 44%, and even to 90% for small and large sample size respectively, when effect sizes are moderate and discrimination is high. The polarized test, which covers the entire range of abilities, albeit very unevenly, shows less dramatic but similarly inflated Type I error rates. Inferential error clearly forms a substantial problem when the true effect consists of two main effects, the test is ill-matched to the sample and items are high in discrimination. Increased sample size only aggravates the situation. The χ^2-test is significant in all these cases.

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

Table 5.2: Number of significant interaction effects (Type I errors) under the additive model with N=120, 360; The expected count is 50

	no effects		1 main effect		2 small effects		2 mod. effects	
N = 120	discrimination		discrimination		discrimination		discrimination	
difficulties	low	high	low	high	low	high	low	high
matched wide	46	55	45	56	53	45	43	55
unmatched wide	55	50	48	45	43	45	72	78
unmatched narrow	59	57	50	53	59	110	120	443
polarized	43	52	56	36	68	68	96	178
$\chi^2(7)$	5.24		6.97		93.41		3772.44	
p-value	0.630		0.432		< 0.0001		< 0.0001	

	no effects		1 main effect		2 small effects		2 mod. effects	
N = 360	discrimination		discrimination		discrimination		discrimination	
difficulties	low	high	low	high	low	high	low	high
matched wide	49	49	49	52	44	51	47	40
unmatched wide	51	47	48	51	50	61	101	142
unmatched narrow	44	50	52	49	82	247	266	900
polarized	55	51	49	42	94	122	258	510
$\chi^2(7)$	1.56		1.68		991.81		21793.56	
p-value	0.980		0.975		< 0.0001		< 0.0001	

5.3.2 Interaction Effects Generated by the Multiplicative Model

For latent abilities generated under the multiplicative model, the number of significant interaction effects corresponds to the power of the statistical test, and is expected to be high. Power is influenced by significance level, sample size, and effect size. Since these last two vary between conditions, the number of significant interactions on the appropriate test (matched wide) was used, averaged for low and high discrimination. This provided a more reasonable expected count. The results for the multiplicative model, including the expected counts and test results[6], are displayed in Table 5.3.

[6] In several cases the expected count for non-significant interactions (not displayed in Table 5.3) 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

Table 5.3: Number of significant interaction effects (power) under the multiplicative model with N=120, 360

	crossing		partial		minimal		orderind	
N = 120	discrimination		discrimination		discrimination		discrimination	
difficulties	low	high	low	high	low	high	low	high
matched wide	999	1000	831	894	845	891	808	871
unmatched wide	996	1000	825	874	793	819	642	696
unmatched narrow	997	983	801	659	727	397	564	85
polarized	998	993	793	749	667	557	470	361
expected count	-		862		868		840	
$\chi^2(7)$	-		554.47		3385.77		7994.55	
p-value	0.010		< 0.0001		< 0.0001		< 0.0001	

	crossing		partial		minimal		orderind	
N = 360	discrimination		discrimination		discrimination		discrimination	
difficulties	low	high	low	high	low	high	low	high
matched wide	1000	1000	1000	1000	1000	999	999	999
unmatched wide	1000	1000	998	1000	996	998	983	993
unmatched narrow	1000	1000	1000	986	993	851	947	167
polarized	1000	1000	995	995	989	942	897	686
expected count	-		-		-		999	
$\chi^2(7)$	-		-		-		804397.40	
p-value	1.000		0.005		< 0.0001		< 0.0001	

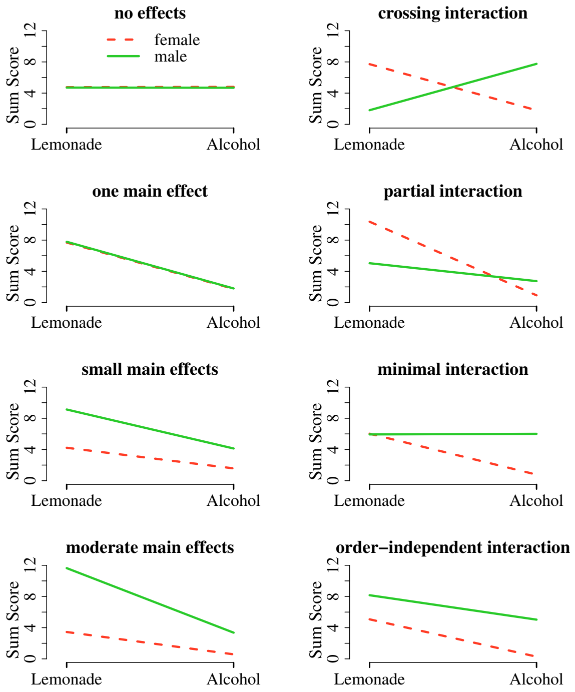
For the crossing interaction, power is extremely high and unaffected by inappropriately targeted, or highly discriminating tests. However, when sample size is small, the other effects all seem to suffer from a decrease in power when the test is more inappropriate and discriminating. Even the partial interaction shows a drop in power. When the test consists of unmatched and narrowly grouped items that are highly discriminating, the power to detect an interaction drops to 0.085. This means 915 out of 1000 interactions

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 marginal as a standard χ^2-test of independence.

were overlooked. This unsettling result is less pronounced in the large sample. For the partial interaction the decrease in power is no longer significant. Of course, this is not unexpected, since an increased sample size improves power. Although the detrimental effect is smaller in the minimal and partial interaction, these findings are nonetheless interesting, since according to Loftus (1978), these types of interactions should be robust against nonlinear monotonic transformation of the scale.

To give an idea of what the observed effects actually look like, the observed means were averaged over iterations and plotted in Figure 5.4. The small and moderate main effects and the order-independent interaction plotted for the observed means obviously look different compared to the effects plotted for the latent means in Figure 5.2.

Figure 5.4: Mean observed effects for unmatched narrow test with high discrimination and sample size 120. Annotated using the alcohol example for illustrative purposes



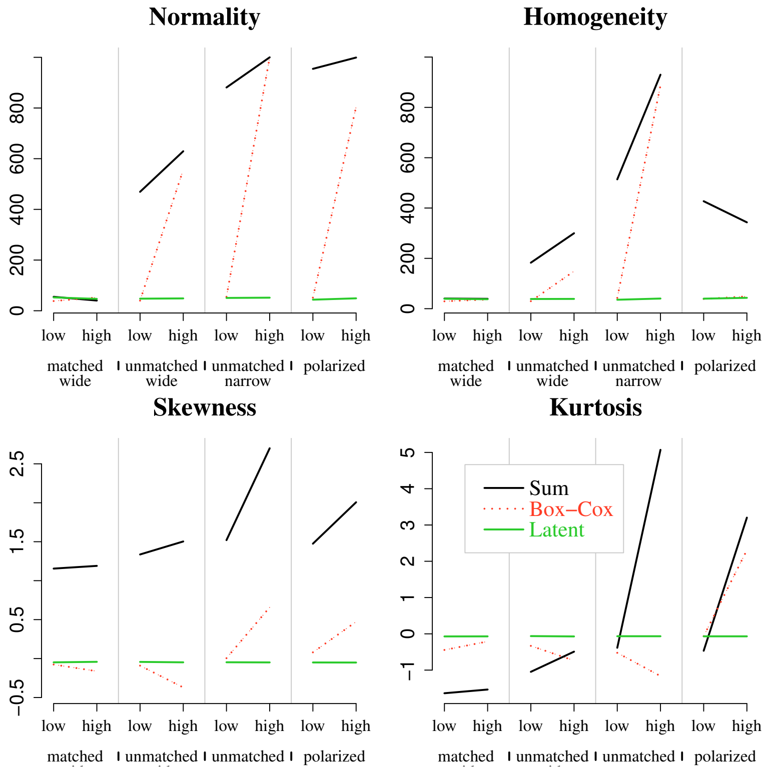
5.3.3 Distributional Shape of Observed and Box-Cox Transformed Scores

It seems highly likely that inferential error is indicated by non-normal or non-symmetric sample distributions. Therefore the assumptions of normality and homogeneity of variances were tested for all combinations of effect type, sample size, discrimination and difficulty. Skewness and kurtosis coefficients were also calculated. Tests and coefficients were determined for both the original observed scores and the Box-Cox transformed observed scores. Figure 5.5 shows the number of significant Shapiro-Wilks and Levene's tests and the mean skewness and kurtosis. The results based on the latent abilities are provided as a base rate. Since the same pattern emerged for different types of effects and sample sizes, the results were collapsed over these factors. Unfortunately this means that tests of normality and homogeneity of variances cannot be used to distinguish between situations that are more at risk for inferential error (e.g. between small and moderate main effects). However, violation of these tests is associated with inappropriate tests in general.

It is interesting to note that virtually all inferential errors could be avoided by considering uninterpretable those effects that coincided with violation of normality or homogeneity of variances. Of course this leads one to err on the side of caution by also discarding a large proportion of valid inferences. Inferential error and violation of normality and homogeneity are not perfectly correlated however. Data are not provided due to limited space, but an informal inspection of the simulated data showed that inferential errors can be associated with samples that do not violate assumptions of normality and homogeneity, but only if the mismatch between items and abilities is less extreme.

As for skewness and kurtosis, the sum scores were skewed to the left with a longer right tail in all cases, but extremely so and moreover highly peaked (leptokurtotic) when discrimination was high and test difficulty was unmatched and narrowly targeted or polarized. As expected, the Box-Cox transformed sum scores show much lower rejection of normality and heterogeneity of variance assumptions and much less extreme skewness and kurtosis.

Figure 5.5: 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



5.3.4 Box-Cox Transformed Scores and Group Differences in Reliability

A final goal of this study was to see whether a normalizing transformation could be used to approximate the latent scores and minimize inferential error. Therefore ANOVAs were also performed on the Box-Cox transformed scores. Tables 5.4 and 5.5 show the main results for the Box-Cox transformed sum scores when true effects were generated using the additive and multiplicative model respectively. The results show the elevated risk of inferential error for tests that are ill-targeted, provide poor scale coverage, and discriminate strongly, can be ameliorated by using the Box-Cox transformation. The Box-Cox transformed scores still show inferential error when there are two moderate main effects, or a partial, minimal or order-independent interaction. However, the number of inferential errors is roughly cut in half compared to those made using the original, untransformed sum scores. Pearson's χ^2-tests were performed on the Box-Cox transformed scores, with results similar to those based on the original scores, although less extreme. Differences in significance as compared to results found on the untransformed scores were found on the single main effect and the minimal interaction. For these effects the number of inferential errors was no longer significantly different from the expected count.

Table 5.4: Number of significant interaction effects (Type I errors) under the additive model for Box-Cox transformed scores with N=120, 360; The expected count is 50

	no effects		1 main effect		2 small effects		2 mod. effects	
N = 120	discrimination		discrimination		discrimination		discrimination	
difficulties	low	high	low	high	low	high	low	high
matched wide	47	56	42	56	54	46	41	49
unmatched wide	60	44	50	52	46	41	51	44
unmatched narrow	57	41	52	53	43	70	53	188
polarized	35	45	55	31	48	52	36	43
$\chi^2(7)$	11.81		10.59		12.34		408.78	
p-value	0.107		0.158		0.090		< 0.0001	

	no effects		1 main effect		2 small effects		2 mod. effects	
N = 360	discrimination		discrimination		discrimination		discrimination	
difficulties	low	high	low	high	low	high	low	high
matched wide	50	49	50	51	46	49	40	40
unmatched wide	54	51	52	49	36	45	53	84
unmatched narrow	45	52	52	53	43	111	52	464
polarized	51	53	45	46	55	52	80	111
$\chi^2(7)$	1.20		1.26		84.99		3734.44	
p-value	0.991		0.989		< 0.0001		< 0.0001	

Table 5.5: Number of significant interaction effects (power) under the multiplicative model for Box-Cox transformed scores with N=120, 360

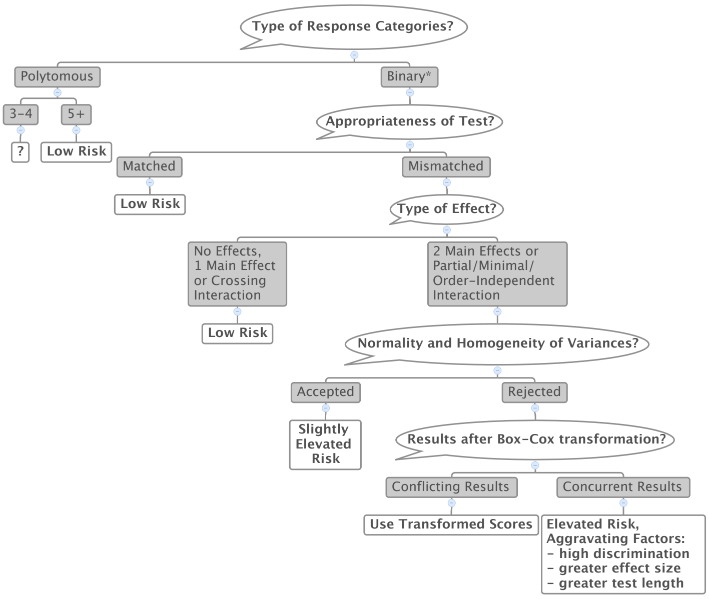
	crossing		partial		minimal		orderind	
N = 120	discrimination		discrimination		discrimination		discrimination	
difficulties	low	high	low	high	low	high	low	high
matched wide	999	1000	825	896	848	899	806	883
unmatched wide	998	1000	826	864	838	885	768	914
unmatched narrow	999	996	807	778	824	649	767	327
polarized	997	1000	800	836	759	782	685	730
expected count	-		860		874		844	
$\chi^2(7)$	-		144.53		704.04		2469.44	
p-value	0.429		< 0.0001		< 0.0001		< 0.0001	

	crossing		partial		minimal		orderind	
N = 360	discrimination		discrimination		discrimination		discrimination	
difficulties	low	high	low	high	low	high	low	high
matched wide	1000	1000	1000	1000	1000	999	1000	999
unmatched wide	1000	1000	998	1000	998	1000	995	1000
unmatched narrow	1000	1000	999	996	998	988	997	751
polarized	1000	1000	996	1000	997	999	982	990
expected count	-		-		-		-	
χ^2(-)	-		-		-		-	
p-value	1.000		0.315		0.118		< 0.0001	

5.4 Assessing Risk

Results from previous studies by Dolan (1994), Loftus (1978), Davison and Sharma (1990), Embretson (1996) and Kang and Waller (2005) can be combined with current results to better identify potential high risk research settings for typical experimental studies. The flowchart displayed in Figure 5.6 provides a decision tree that addresses the most important factors in an efficient order. Assuming the dependent variable is measured with a questionnaire or test, the first consideration is the number of response categories. When items are polytomous, with five to seven or more response categories,

Figure 5.6: Identification of risk factors for inferential error, assuming a quantitative latent variable and a logistic link-function



the relation between latent and observed scores is assumed to approximate a linear function and risk of inferential error is low[7]. For three or four categories the situation is less clear, but when items are binary, an elevated risk of inferential error is a real possibility. However, if we can assume the relation between latent and observed scores is accurately described by a Rasch or two-parameter IRT model, a further assessment can be made.

If such an assumption is reasonable, the next consideration is whether the test is appropriate for the sample. If items are well matched to the sample,

[7] Even if we assume the probabilities of choosing each response category are accurately described by logistic functions as in the IRT Graded Response Model, the aggregate test response curve will be close to linear. Preliminary findings show this is also true for items that have three to five response categories (Zand Scholten & Borsboom, 2010b). The situation for three to five response categories has not been assessed for other types of models or functions however.

risk of inferential error is low. However, when the items are very difficult or very easy and provide poor coverage of the ability scale, inferential error could occur (Kang & Waller, 2005). If this is the case, the next step is to consider the type of effect, since not all effects are equally vulnerable to inferential error.

If the group means show no effects, one main effect or a crossing interaction, risk of inferential error is low. In contrast, if the group means show an order-independent interaction or two main effects, caution is warranted (Loftus, 1978). Due to the combined effect of nonlinear transformation and sampling error, the same applies to partial and minimal interactions, albeit to a lesser extent. Fortunately, when tests show that normality and homogeneity of variances can be accepted, risk of inferential error is only slightly elevated. However, when normality or homogeneity have to be rejected, this could be due to a distorting effect of the nonlinear relation. Scores can be transformed to normalize the data, by using a procedure such as the Box-Cox transformation.

When results from an analysis based on Box-Cox transformed scores oppose those based on the original scores, the transformed scores can be used as an approximation of the latent ability values, similar to latent ability estimates obtained from IRT modeling. When results concur with those based on the original scores, we have to accept that the risk of inferential error is still present, although its magnitude is hard to assess. Some factors that can be considered are effect size, test length, item discrimination and sample size. Effect size and test length increase the risk of inferential error as they grow larger. For sample size, the same applies if the true effect is additive, but if a true interaction is present, a larger sample size is beneficial and decreases the risk of inferential error. Of course, in practice, the true effect is unknown, so that the impact of sample size is hard to determine. An effect that is more readily interpretable is the effect of item discrimination. For almost all effects low item discrimination can greatly ameliorate the risk of inferential error. A test with moderate item discrimination is only at risk for inferential error when effect size is large and the effect consists of an order-independent interaction or two main effects.

5.5 Discussion

Our results show that there is a substantial risk of making inferential errors concerning interaction effects under conditions that are typical for many experimental studies in psychology. This danger is only clear and present however, when a test with dichotomous items is employed that is inappropriately difficult (or easy). Under these conditions, the number of spurious interactions or Type I errors increased sharply when the true effects consisted of two moderate main effects. Increasing item discrimination, which corresponds to improving the reliability of the test, only exacerbated the problem. The situation was even more serious when interactions were actually present on the latent level. More than 90 percent of the latent order-independent interactions was overlooked when observed scores on an ill-targeted test with poor ability scale coverage and high discrimination. Of course we chose an extremely difficult test to better illustrate the problem, but it is apparent that mismatched tests can cause serious difficulties for experimental researchers.

The rejection rate of the tests of normality and homogeneity of variances increased when the test was inappropriate for the sample. This result is not unexpected. When a test is inappropriate, group means will tend to be more extreme, resulting in ceiling or floor effects that are associated with skewed sample distributions and reduced variance. The skewness and kurtosis coefficients underlined this. Unfortunately, these results are uniformly found for all types of effects. This means rejection of normality and equal variance assumptions is linked to test inappropriateness, but cannot be used to differentiate between situations that are less or more at risk for inferential error.

Our results replicate earlier findings that suggest appropriate coverage of the ability sample is essential in guarding against invalid inference concerning interaction effects. This means that where possible, tests need to be psychometrically sound, consisting of a moderate number of items, with moderate item discrimination and item difficulties that are targeted at the mean ability that is expected in the sample. In other words, items need to be selected that produce a test characteristic curve that is maximally linear. It is interesting to note that actions to improve the internal and external validity of research findings that are generally considered good practice, such as increasing effect size, sample size, and test length, can be highly counterproductive when it comes to mitigating the risk of inferential errors.

But what if we are unable to use appropriate tests? Applying a Box-Cox transformation to the observed scores resulted in less Type I errors under the additive model and higher power to detect significant interactions under the multiplicative model. This transformation also improves rejection rates for tests of normality and homogeneity of variances. Skewness and kurtosis are also much less extreme, which is to be expected of a transformation that is aimed at normalizing the data. Thus it seems that a Box-Cox transformation can be used to kill two birds with one stone. Both statistical test assumptions of normality and equal variances will be met more often, and the risk of inferential errors is minimized.

In fact, the Box-Cox transformation can be viewed as an attempt to approximate the latent values, which we assume to be distributed normally, without actually having to model the latent variable. Of course IRT modeling or generalized linear mixed modeling could provide a more elegant solution to the problem of inferential error raised here, but since such methods are not always available it is important to have an alternative way of dealing with the risk of inferential error.

We have narrowed down potential problem situations to settings with inappropriate tests that use dichotomous items, showing either minimal, partial or order-independent interactions, or two main effects. This reduction comes at the cost of several additional assumptions. The most important of these assumptions concerns the specification of the relation between the latent and observed variable. We have chosen the two-parameter model because it is simple and has the monotone likelihood ratio property, which entails that the items result in a consistent ordering of persons abilities based on the sum score. Although these models perhaps form an oversimplified representation of the relation between latent and observed level, some sort of logistic-type function seems plausible. Other models can be considered that introduce more parameters that model characteristics such as guessing or slippage, for example. The effect of additional parameters on the linearity of the test response curve could provide more insight into our understanding of the risk of inferential error.