The choice of product indicators in latent variable interaction models: post hoc analyses

Njål Foldnes
BI Norwegian Business School

Knut Arne Hagtvet
University of Oslo

This is the authors' accepted and refereed manuscript to the article published in

***Psychological Methods,* 19(2014)3:444-457**

DOI: http://dx.doi.org/10.1037/a0035728

The Choice of Product Indicators in Latent Variable Interaction Models: Post Hoc Analyses

Njål Foldnes
BI Norwegian Business School

Knut Arne Hagtvet
University of Oslo

Author Note

The authors would like to thank Jostein Rise and his coworkers Silje Sommer Hukkelberg and Velibor Bobo Kovac at The Norwegian Institute for Alcohol and Drug Research for providing us with data.

Correspondence concerning this article should be addressed to *njal.foldnes@bi.no*

Abstract

The unconstrained product indicator (PI) approach is a simple and popular approach for modeling nonlinear effects among latent variables. This approach leaves the practitioner to choose the PIs to be included in the model, introducing arbitrariness into the modeling. In contrast to previous Monte Carlo studies, we evaluated the PI approach by three post-hoc analyses applied to a real-world case adopted from a research effort in social psychology. The measurement design applied three and four indicators for the two latent first-order variables, leaving the researcher with a choice among more than 4000 possible PI configurations. Sixty so-called matched-pair configurations that have been recommended in previous literature are of special interest. In the first post-hoc analysis we estimated the interaction effect for all PI configurations, keeping the real-world sample fixed. The estimated interaction effect was substantially affected by the choice of PIs, also across matched-pair configurations. Subsequently, a post-hoc Monte Carlo study was conducted, with varying sample sizes and data distributions. Convergence, bias, type I error and power of the interaction test were investigated for each matched-pair configuration and the all-pairs configuration. Variation in estimates across matched-pair configurations for a typical sample was substantial. The choice of specific configuration significantly affected convergence and the interaction test's outcome. The all-pairs configuration performed overall better than the matched-pair configurations. A further advantage of the all-pairs over the matched-pairs approach is its unambiguity. The final study evaluates the all-pairs configuration for small sample sizes, and compares it to the non-PI approach of LMS.

Keywords: latent interaction, product indicators, unconstrained approach, post hoc analysis, matched-pair strategy

The Choice of Product Indicators in Latent Variable Interaction Models: Post Hoc Analyses

Introduction

Models with nonlinear relationships among latent variables are often encountered in social and behavioral sciences. In this paper we investigate an instance of the classical Kenny and Judd (1984) model where two latent predictor variables ξ_1 and ξ_2 interact to have a nonlinear effect on a latent criterion variable η:

$$\eta = \gamma_1\xi_1 + \gamma_2\xi_2 + \gamma_3\xi_1\xi_2 + \zeta. \tag{1}$$

In the product indicator (PI) approach to estimating (1), the latent product term $\xi_1\xi_2$ is represented by some set or configuration of PIs x_ix_j, where x_i and x_j are indicator variables of ξ_1 and ξ_2, respectively. It is up to the researcher to decide which PIs to include. Prior research and practice have applied from one (the 1-pair approach) to all possible PIs (the all-pairs approach). No clear consensus yet exists concerning the type and number of product indicators to be included in the model to operationally define the latent product term. This choice is largely left to the applied researcher, introducing arbitrariness into the modeling process. However, in an influential study Marsh, Wen, and Hau (2004) recommended a *matched-pairs* strategy that greatly reduces the number of possible PI configurations. This strategy is based on two general suggestions: a) use all information, that is, all of the first-order indicators should be used in the formation of PIs, and b) do not reuse any of the information; that is, each of the first-order indicators should be used in only one PI. In Marsh et al. (2004), ξ_1 and ξ_2 each had three indicators, and the suggested 3-match strategy implied three PIs in favor of all the nine PIs in their study. The majority of possible PI configurations violate either the use-all-information or the do-not-reuse-information principles. For instance, the 1-pair strategy violates the first principle, whereas the all-pair strategy violates the second principle. The matched-pairs strategy implies that an intermediate position has to be taken where the researcher is left

with a choice somewhere between one or all possible PIs. However, the problem of arbitrariness remains to some degree, as there are different PI configurations that all adhere to the matched-pair strategy. For instance, if each of ξ_1 and ξ_2 has three first-order indicators, there are six matched-pair configurations, one of which must be chosen for interaction modeling.

The present study is concerned with whether variation across possible PI configurations might have an impact on inference regarding the interaction effect. It is obvious from statistical theory that two different PI configurations will lead to differences in parameter estimation and model evaluation. However, in a typical real-world sample, what is the extent of this difference? Is the estimated interaction effect substantially affected by the choice of PIs? If so, is the variation reduced when limiting ourselves to matched-pair configurations? If substantial variation is found across configurations, also within the matched-pair strategy, which approach to forming PIs is preferable in terms of convergence rates, bias, type I error and power to detect an interaction effect? Any recommended approach should reduce ambiguity by giving clear-cut advice to applied researchers.

To investigate these questions, three post hoc studies based on a real-world sample were conducted. Basing the investigation on a real-world sample contrasts with the Monte Carlo methodology underlying most of our current knowledge about the performance of various latent interaction modeling strategies (e.g. Wall & Amemiya, 2001; Marsh et al., 2004; Little, Bovaird, & Widaman, 2006; Klein & Muthén, 2007). In these studies, the researcher has complete knowledge of the underlying population structure and is free to manipulate design variables like sample size, deviation from non-normality, and level of misspecification. If the simulated conditions differ from those found in real-world data, the conclusions do not carry over to the researcher's real-world situation. By instead focusing on a real-world sample we study a situation which more likely reflects the complexity of real-world data. The generalizability of our results will however be limited to situations that are similar to our empirical example.

In study 1 the real-world sample is held fixed, and we focus on different ways to operationally define the latent interaction variable in terms of number and organization of the PIs given the number of available choices in the actual measurement design. This results in a variation across more than 4000 PI configurations. Study 1 examines the variation in estimated interaction effect γ_3 across the large number of PI configurations.

Study 2 is a post hoc Monte Carlo study based on the same real-world data. In this study we examined the variation across simulated samples and across 61 selected PI configurations: In line with prior research we investigated all 60 different versions of the matched-pairs strategy plus the all-possible pairs strategy. The variation across matched-pair configurations for a typical sample was investigated. We also provide information for each matched configuration and the all-pairs configuration in terms of convergence, bias, coverage, and type 1 error and power for testing the interaction effect.

Based on study 2 we recommend the all-pairs configuration when using PIs to model latent interaction. The currently most popular alternative to PIs is the latent moderated structural equations (LMS) approach (Klein & Moosbrugger, 2000). In study 3 we perform a post hoc Monte Carlo comparison of all-pairs and LMS under realistic conditions, i.e. with small sample sizes and non-normal data.

While the post hoc Monte Carlo simulations in studies 2 and 3 deviate from typical a priori Monte Carlo studies, our post hoc simulation across models in study 1 is rarely done and we are not familiar with any such study for the purpose of the present study. The present combination of the three post hoc studies are not either typical in this research area. This combination allows assessing to what extent findings are generalizable across the post hoc studies.

This article is organized as follows. We first present the PI approach to interaction modeling. Next we present our real-word case from social psychology. This is followed by the three studies. Each study is reported by the three sections; method, results and discussion. The article is ended by a general discussion and a conclusion.

The PI Approach

The PI approach as presented in the seminal paper of Kenny and Judd (1984) originally included a set of parameter constraints that were complicated to incorporate in the model. Specifying such nonlinear constraints on the model parameters has proved to be challenging for applied researchers, and has not resulted in much applied work. Another limitation is that the parameter constraints are deduced under the assumption of multivariate normal data, a condition that is often violated. A partially constrained approach, where ξ_1 and ξ_2 are not assumed to be jointly normal, was proposed by Wall and Amemiya (2001) and was found to outperform the constrained approach in situations where the factors were non-normally distributed.

The complexity of the nonlinear constraints and their dependency on the normality assumption led Marsh et al. (2004) to abandon these constraints. Although the unconstrained approach is easier to implement than are constrained approaches, the researcher still must choose which PIs to use as indicators for $\xi_1\xi_2$. In published evaluations of the PI approach, the number of PIs have varied from all possible pairs (Kenny & Judd, 1984; Wall & Amemiya, 2001; Marsh et al., 2004; Yang-Wallentin, Schmidt, Davidov, & Bamberg, 2004; Little et al., 2006; Steinmetz, Davidov, & Schmidt, 2011) to a reduced number of indicators (Jöreskog & Yang, 1996; Yang-Wallentin, 1998; Wall & Amemiya, 2001; Marsh et al., 2004; Saris, Batista-Foguet, & Coenders, 2007). In the all-possible-pairs strategy, all combinations $x_i x_j$, where x_i is an indicator of ξ_1 and x_j is an indicator of ξ_2, are used to represent the latent product term.

Monte Carlo results by Wall and Amemiya (2001) indicate that the all-pairs strategy seems preferable to the one-pair strategy. This was reiterated by Marsh et al. (2004), which finds that the sampling fluctuations of the estimated interaction effect γ_3 was larger for the one-pair strategy. The all-pairs strategy was also employed by Little et al. (2006). Marsh et al. (2004) remarked that with one pair, only a small part of the available data is used, while with the all-possible-product strategy one repeatedly reuses the same information in

terms of single indicators for the main-effect factors. They argue for a matched-pair strategy where all information is used, but not repeatedly. By this they mean that each indicator of ξ_1 and ξ_2 should appear as a constituent component of a PI only once. They contend that using an indicator more than once to form PIs may introduce non-parsimonious reuse of the information contained in the indicator.

With the exception of a few studies (Marsh et al., 2004; Wall & Amemiya, 2001) the effect of varying the number of PIs for the same latent interaction variable does not seem to have been examined. Marsh et al. (2004) compare latent interaction models with one, three, and nine PIs, respectively. Based on their Monte Carlo studies, they suggest a reduced number of indicators in terms of their 3-match configuration in favor of the 9-product configuration (i.e., all-pairs configuration). Their preference for a more parsimonious model was based upon the observation that the more complicated 9-pair configuration does not display substantial improvement compared to the 3-match configuration. Nevertheless, Marsh et al. (2004) and Marsh, Wen, and Hau (2007) recommend further study to find an optimal strategy to construct PI configurations. For instance, when using an unequal number of indicators, the principle of "do not reuse information" may not easily be applied.

To develop a well-specified interaction model requires some assumptions in the measurement model. The central assumption of normally distributed data that underlies the constrained PI approach is unnecessary in the unconstrained approach. However, model formulation will be based on the following assumptions:

1. If two indicator residuals (e.g., δ_1 and δ_2) are uncorrelated, they are also independent.
2. δ_i and ξ_j are independent for all pairs i, j.
3. All residuals and the main factors ξ_1 and ξ_2 have expectation zero.

These assumptions are needed to establish whether correlated uniquenesses are called for when introducing PIs into the model. As can be shown by covariance algebra, they imply

that the residuals of a first-order indicator and a PI are uncorrelated. That is, if δ_{ij} and δ_i are the residuals of x_{ij} and x_i, respectively, then δ_{ij} and δ_i are uncorrelated. However, we need to model the covariance between the residuals of two PIs whenever they share a constituent indicator variable. For instance if both x_1x_4 and x_2x_4 are included their uniquenesses will be correlated because of the common indicator x_4.

An additional complication that, to the best of our knowledge, has not appeared in previous PI literature is the modeling of covariance between the residuals for the same main-effect indicator. It is sometimes necessary for conceptual reasons to include residual covariance between to first-order indicators for the same latent construct. Suppose for instance that x_1 and x_2 are indicators for ξ_1 while x_3 and x_4 are indicators for ξ_2. If there is a conceptually based residual covariance between x_1 and x_2, then the residual covariance between x_1x_3 and x_2x_4 must be modeled.

Finally, we note that there are alternative approaches to estimating the interaction model (1) that do not rely on PIs, e.g., the LMS approach and the method of moments approach (Mooijaart & Bentler, 2010). These recent methods are based on more elegant underlying theory than the PI approach. However, more research is needed to evaluate their performance under realistic conditions, i.e. with non-normal data and finite sample sizes. In study 3 with compare the LMS approach with the all-pairs configuration under such conditions.

Assessing an Interaction Hypothesis in Social Psychology

Our point of departure was the study by Hukkelberg, Hagtvet, and Kovac (2013) where an interaction hypothesis was investigated involving three constructs: attitude (ATT; positive or negative evaluations of a behavior), perceived behavioral control (PBC; the perception of whether performing the behavior is achievable) and the dependent variable goal commitment (GC; the intention to perform a given behavior). It was hypothesized that for individuals who are positive about quitting smoking, the relationship between

PBC and GC (the intention to quit smoking) would be strong, while the same relationship would be weak or non-existing for individuals who are negative about quitting smoking.

PBC was measured with three indicators p_1, p_2, and p_3. ATT was measured by four indicators a_1, a_2, a_3, and a_4. GC was measured by three indicators g_1, g_2, and g_3. Further details and definitions of the indicator variables are described in the Appendix. To exemplify, consider the interaction model based on the following configuration of PIs: $p_1a_1, p_1a_2, p_2a_3, p_3a_4$. This model is in line with the matched-pair strategy dictated by the "use-all-information" principle. The path diagram of the interaction model based on this set of PIs is given in Figure 1. Two correlated uniquenesses for the PIs were necessary.

The sample

The data in this study derive from a longitudinal study conducted in November 2006 and March 2007 on smoking. Smokers, aged 15 to 74, were invited to respond to the questionnaire through an invitation displayed in 15 internet newspapers over a 10-day period. Altogether, we had access to 939 daily smokers who responded to the initial invitation and to the follow-up questionnaire. Data were analyzed using complete case analysis, resulting in a sample size of $n = 926$ participants.

The correlation matrix and descriptive statistics for the 10 indicator variables are given in Table 1. The marginal distributions are all highly non-normal, as is also confirmed by the Jarque-Bera test of normality.

The Measurement Model

We first fit the measurement model for the three latent variables, omitting the latent product term. This approach allowed us to check whether the fit was adequate before we proceeded to the interaction model. Moreover, testing the measurement model allows testing for any residual covariances among the main-effect indicators. In our sample, we had substantive reasons to expect correlation between residuals for indicators of the latent attitude variable. As described in the Appendix, the indicators a_1 and a_2 are cognitively

oriented, while indicators a_3 and a_4 are affect indicators of the same latent attitude variable. However, due to the distinction between the two types of indicators, correlations between indicators within each type may be expected because they may share a unique factor over and beyond the common attitude factor. The fitting process of the measurement model required a correlation between the residuals for the indicators a_3 and a_4. The measurement model with standardized estimates is given in Figure 2.

Normal theory based maximum likelihood (ML) estimation was used to estimate the model. Because our data are non-normal the model fit chi-square is that of Satorra and Bentler (1994) and reported fit indices like the root mean squared error of approximation (RMSEA) and comparative fit index (CFI) are based on the Satorra-Bentler chi-square. Due to non-normality, distribution-robust standard error estimation was employed (Satorra & Bentler, 1994, equation 16.10). Overall, we deem the measurement model to have reasonable fit: $\chi^2(31) = 192.3$, RMSEA=.075, CFI=.96, SRMR=.068.

Study 1: Post Hoc Analysis for the Real-World Sample

Method

A central goal in this study was to investigate how the estimated interaction effect depends on the choice of PIs. For each PI configuration the estimated effect will be different, and therefore the question is how much the effect varies between configurations. We excluded configurations with only one PI because these do not yield an identified model under the unconstrained approach. The total number of possible configurations is then $2^{12} - 12 - 1 = 4083$. Figure 1 depicts one of these 4083 configurations. In study 1 we investigate the variation among configurations in terms of proper solutions and the estimated value and significance of the interaction parameter.

To compare the estimated interaction γ_3 effect across models, we use its standardized value. As pointed out by Friedrich (1982), in multiple regression the standardized estimate $\dot{\gamma}_3$ of γ_3 is incorrect. This carries over to the standardized estimate reported in SEM

software packages, with the correct formula given by Wen, Marsh, and Hau (2010):

$$\tilde{\gamma}_3 = \dot{\gamma}_3 \sqrt{\frac{\widehat{var(\xi_1)}\widehat{var(\xi_2)}}{\widehat{var(\xi_1\xi_2)}}}. \tag{2}$$

A main concern in interaction models is to establish whether the interaction is significant, that is, whether the parameter γ_3 is significant. Significance testing of γ_3 might use the z-value of γ_3, $z = \widehat{\gamma}_3/s.e.(\widehat{\gamma}_3)$, or alternatively employ a scaled χ^2 difference test. In this paper, we employ the z-value test based on robust standard errors.

The current study especially investigates variation within the matched-pair strategy. Due to the unequal number of indicators for PBC and ATT there is a conflict between the two principles of Marsh et al. (2004) to a) use all information, and b) do not reuse any of the information. If principle a) is given precedence then we would need four PIs to match up all the four indicators of ATT, while if principle b) is deemed more important we should only use three PIs, one for each indicator of PBC. For the latter case, we define the class 3MATCH as containing all matched-pair configurations with three product indicators. In this class each of p_1, p_2 and p_3 appears in exactly one PI, while no a_j appears in two PIs. For instance the configuration p_1a_1, p_2a_2, p_3a_4 belongs to 3MATCH. The total number of PI configurations in 3MATCH is 24. For the former case, we define the class 4MATCH as containing all PI configurations in which each of the a_j is used in exactly one PI, while each of the p_i is used once or twice. The configuration in Figure 1 belongs to 4MATCH. There is a total of 36 configurations in 4MATCH. The all-pairs configuration consisting of all twelve PIs is denoted by ALL.

All models contained correlated uniquenesses between pairs of PIs with a common constituent indicator, and between p_ia_3 and p_ja_4 whenever these were included, as exemplified in Figure 1. The product term $\xi_1\xi_2$ was scaled by fixing the loading of the first PI, say x_ix_j, in the configuration to 1.0, where the ordering of the PIs was $p_1a_1, p_1a_2, \ldots, p_3a_3, p_3a_4$. The latent variables ξ_1 and ξ_2 were then scaled by fixing the loadings of indicators x_i and x_j to 1.0. Our sample was double-mean-centered, that is, we

first centered each of the observed variables, and then formed the PIs and recentered them. As shown by Lin, Wen, Marsh, and Lin (2010), such double-mean-centering of the data allows dropping a mean structure in the model. Each of the 4083 models was estimated with ML using the R package lavaan (Rosseel, 2012). Robust standard errors were used, because of the non-normality of the data. The variables are ordinal, violating the continuity assumption underlying ML estimation. However, as variables are measured on a 7-point scale, the ML method is acceptable (Rhemtulla, Brosseau-Liard, & Savalei, 2012).

Results

Improper Solutions. A configuration obtained a fully proper solution if the estimation converged to an admissible estimate. Improper solutions (i.e. non-convergence and Heywood cases) were excluded in the investigation of γ_3 and its z-value. We obtained fully proper solutions for 85% of the 4083 models. The likelihood of obtaining a proper solution increases with an increasing number of indicators, as shown in Table 2. So from a convergence viewpoint, the more PIs the better. Among the matched-pair configurations we obtained fully proper solutions in 22 of the 24 configurations in 3MATCH, and in 35 of the 36 configurations in 4MATCH.

Variation of Estimated $\tilde{\gamma}_3$ across Models. The distributions of $\tilde{\gamma}_3$ is given in the kernel density plot in Figure 3. The main observation drawn from Figure 3 is that there is a large variation of the estimated interaction effect across all PI configurations. The mean and standard deviation of $\tilde{\gamma}_3$ were .134 and .107, respectively. Hence, varying the PI configuration significantly affects the estimated interaction effect.

Also within 3MATCH and 4MATCH we found considerable variation in $\tilde{\gamma}_3$, as shown in Figure 4. We observe that for the 22 configurations in 3MATCH there is considerable variation, from p_1a_1, p_2a_3, p_3a_2 resulting in the minimal value $\tilde{\gamma}_3 = .05$, to the maximum value of $\tilde{\gamma}_3 = .29$ obtained with p_1a_3, p_2a_2, p_3a_4. Although the variation in $\tilde{\gamma}_3$ within 4MATCH was less than within 3MATCH, the 35 configurations in 4MATCH still resulted

in notable variation, from the minimal $\tilde{\gamma}_3 = .04$ obtained with $p_1a_1, p_2a_3, p_3a_2, p_3a_4$, to the maximal $\tilde{\gamma}_3 = .21$ obtained with $p_1a_3, p_2a_1, p_2a_3, p_3a_4$. The ALL configuration resulted in $\tilde{\gamma}_3 = .12$, and is depicted for reference in Figure 4.

Variation of the Significance of γ_3 across Models. The significance of γ_3 was evaluated at the $\alpha = .05$ level, so the decision rule is to reject $H_0 : \gamma_3 = 0$ if $|z|$ exceeds 1.96. The null was rejected in 3360 of the 3487 models, i.e. 96 % of the PI configurations resulted in a significant interaction effect. The z-value varied markedly across PI configurations, with mean and standard deviation 4.06 and 1.00, respectively. The distribution of z is depicted in the kernel density plot in Figure 5.

The variation of z within the matched-pair configurations in 3MATCH and 4MATCH is represented in the boxplots in Figure 6. In line with the findings for $\tilde{\gamma}_3$ the variation is notable. In 3MATCH there is one configuration, namely p_1a_1, p_2a_3, p_3a_2, that leads to a non-significant interaction effect. For 4MATCH there are two configurations that results in a non-significant z, namely $p_1a_1, p_2a_3, p_3a_2, p_3a_4$ and $p_1a_1, p_1a_4, p_2a_3, p_3a_2$. There is larger variation among the 3MATCH configuration compared to 4MATCH. The ALL configuration yielded a z-value of 4.23, and is included in Figure 6.

Discussion

It was expected that the statistical significance of the interaction parameter would depend on the product indicators used in the model. However, the extent of this variation is not easy to establish a priori. To our knowledge, no exhaustive analysis of the variation across models has been conducted previously in the empirical and theoretical literature. In the present investigation we found considerable variation in terms of proper solutions, parameter estimates and significance. Hence the modeling outcome for an applied researcher choosing a specific configuration is heavily dependent upon this choice. Existing literature suggests using only a subset of the PIs according the the principle of including every first-order indicator only once in a PI. This approach is reasonable and fortunately

limits the choice of possible PI configurations substantially. However, there is still ample room for a researcher to choose within this limited class. We found that the variation within the MATCH classes was substantial. For instance, under some matched-pair configurations the interaction was non-significant while it was highly significant for most other configurations. So the ambiguity inherent in the PI approach is still a problem of practical importance even under the matched-pair strategy.

Study 2: Monte Carlo Extension Based on the Real-World Sample

Method

Study 1 investigated variation in estimated interaction effect across 4083 PI configurations, keeping the single real-world sample fixed. In Study 2 we complement this analysis by Monte Carlo simulation of artificially generated samples. We retain 61 configurations of special interest to our research questions, namely all 24 members of 3MATCH, all 36 members of 4MATCH, in addition to the all-pairs configuration. We investigate how convergence, standard error bias, type I error and power to detect interaction vary among three design conditions: the presence/absence of interaction effect in the data, sample size and distribution of the data. These are typical design parameters in conventional Monte Carlo studies. In such studies it is typically the case that only a handful of models are evaluated, representing different model complexities. In study 2, however, variation across models is a major issue, with 61 PI configurations being evaluated. The performance of each configuration under varying design conditions is obtained by aggregating over all replications in each condition. However, this information does not fully answer the question of how large the variability across matched-pair configurations is in a typical sample. This is of essential concern for a researcher wishing to model interaction based on a single sample. If the matched-pair strategy is to be used, what might the consequences be of choosing one matched-pair configuration over another? Does the choice matter? To obtain information about the variability across candidate

configurations, we calculated in each sample the range of $\tilde{\gamma}_3$ and z. The variation in these statistics across matched configurations in a typical sample is estimated by taking the mean over all replications in each of the eight design cells.

In accordance with the post hoc principle adopted in this paper, we randomly draw artificial samples from a model similar to the estimated interaction model obtained from the real-world sample described in study 1. That is, we fix the free parameters in the interaction model to values close to those estimated with the real-world sample. Consequently, the simulated samples have characteristics that resemble the real-world sample. Further in line with the post hoc approach the Monte Carlo design includes conditions that are close to those found in our real-world sample. For each design parameter we define one condition to match the real-world sample, and one contrasting condition. The design parameters are sample size, distribution and interaction effect:

- Real-world interaction effect versus no interaction effect in data.
- Real-world non-normal distribution versus multivariate normal distribution.
- Real-world sample size $n = 926$ versus small sample size $n = 200$.

The full factorial design yields $2 \cdot 2 \cdot 2 = 8$ conditions, with 1000 sample replications in each condition. Sixty one models were estimated for each sample in each condition, totaling 488000 model estimations. Detailed description of all 61 configurations included in study 2 is available in online supplementary materials (Table 3). The 24 configurations in 3MATCH were named $3M_1$, $3M_2$, ..., $3M_{24}$. Similarly the 36 4MATCH configurations are denoted by $4M_1$, $4M_2$, ..., $4M_{36}$. The all-pairs configuration is denoted by ALL. We next describe the details of the data generation.

Each random sample was generated by first generating values for the latent variables PBC and ATT with variances and covariance equal to the values obtained when estimating the measurement model in Figure 2. Then values of p_i and a_j were generated by applying loading coefficient and error term variances/covariances also obtained from the

measurement model. Then values for the latent construct GC were generated by

$$GC = 0.4 \cdot PBC + 0.4 \cdot ATT + \gamma_3 PBC \cdot ATT + \zeta$$

where the regression coefficients and the variance of ζ are approximately equal to the values obtained from estimating the interaction model using the real-world sample with the all-pairs configuration. In the condition where an interaction effect was present, we set $\gamma_3 = 0.15$, which is close to the value obtained with the all-pair configuration in the real-world sample. Under the assumption of multivariate normality, the interaction effect size accounted for 2.95% of $var(GC)$, and the squared multiple correlation was $R^2 = .246$.

Data without interaction effect were obtained by fixing $\gamma_3 = 0$. Finally, the observed values g_j were obtained by applying loading coefficient and error term variances from the estimated measurement model.

Distributional characteristics similar to those observed in the real-world sample were obtained by setting the skewness of PBC and ATT to -0.5 and -1, and excess kurtosis to 2 and 10, respectively. All other random constituents, i.e. residual error terms and ζ, were normally distributed. Table 4 contains large-sample estimates of indicator skewness and kurtosis in the non-normal Monte Carlo condition. The values, although not identical, are reasonably comparable to the corresponding values observed in the real-world sample in Table 1.

Data generation and estimation were done in R and in the lavaan package. Non-normal data were generated from the Johnson distribution as implemented in the R SuppDists package. The generated samples were double-mean centered and the z-value and confidence interval were obtained from the ML estimator with distribution-robust standard errors.

Results

Variation across matched-pair configurations in a typical sample. In a practical situation a researcher has only one sample at hand. It is of interest to find out

how much variation there is across 3MATCH and 4MATCH in a typical sample. To illustrate how a statistic derived from different configurations might vary on a given single sample, we calculated the range of $\tilde{\gamma}_3$ and of z across the 3MATCH- and across the 4MATCH configurations for each replicated sample. That is, in each replicated sample, the max and min value of of $\tilde{\gamma}_3$ and z across matched-pair configurations was calculated. The mean of these values for $\tilde{\gamma}_3$ and z across all (close to 1000) replications in each of the eight design cells are tabulated in Tables 5 and 6, respectively.

In all design conditions, there is considerable variation in both $\tilde{\gamma}_3$ and z across configurations. Despite the fact that 4MATCH contains 12 more configurations than does 3MATCH, the mean range is smaller across 4MATCH configurations than it is across 3MATCH configurations, in all eight conditions, for both $\tilde{\gamma}_3$ and z. For both $\tilde{\gamma}_3$ and z, the range is larger when interaction is present. Also, with non-normal data the overall range is slightly larger than for normal data. With larger sample size the range of $\tilde{\gamma}_3$ decreases. The same effect is found for z, provided there is no interaction in the data. For data with interaction, the range of z is larger for the large sample size.

Next we evaluate the performance of each of the 61 models separately, by aggregating the performance of each model over replications.

Convergence. Nonconvergence occurs when the model-implied covariance matrix at some point is no longer positive-definite. In most cases, nonconvergence is caused by Heywood cases, where some variance is estimated to be negative. The percentages of converged solutions for each model and in each of the eight design cells can be found in online supplementary materials (Table 7).

The overall convergence rate across all configurations and design conditions was 94%. A design factor with strong impact on convergence concerned whether an interaction effect was present in the data. Aggregating over all configurations, sample sizes, and data distributions, the convergence rate for data with interaction ($\gamma_3 = .15$) was 97%, while for data without interaction ($\gamma_3 = 0$) the rate was 91%. Hence, the presence of an interaction

effect raised the likelihood of obtaining a converged solution. In the absence of an interaction effect in the data, the interaction model is still correctly specified, but the regression of η on $\xi_1\xi_2$ is superfluous. This over-fitting might explain the decrease in convergence for data with no interaction effect.

Aggregating over all configurations, sample sizes, and presence/absence of interaction, the convergence rates were overall slightly higher for the non-normal distribution, namely 94.5% under non-normality and 93.5% with normal data.

As expected, sample size affected convergence rate. Over all configurations, data distributions and presence/abscence of interaction, convergence rate increased from 92% to 96% for sample sizes $n = 200$ and $n = 926$, respectively.

Next we consider variation within the PI configurations, aggregating over the eight design conditions. Convergence rates were lower in 3MATCH than in 4MATCH, with respective overall convergence rates being 90.1% and 96.5%. The all-pairs configuration ALL has an overall convergence rate of 99.9%, higher than the convergence rate of all the 60 matched-pair configurations. Hence, we found no support for the concern raised by Marsh et al. (2004) that increasing the number of PIs might lead to nonconvergence.

Convergence rates varied markedly between the matched-pair configurations. Within 3MATCH convergence rates across all eight conditions varied from 76.8% (3M18) to 99.7% (3M11), while for 4MATCH convergence rates varied from 81.9% to 99.8%.

The discussion in the following sections is based on converged solutions.

Parameter bias, efficiency and coverage. To compare the precision and efficiency in estimating γ_3 across models we consider several performance criteria. For conditions with no interaction effect ($\gamma_3 = 0$), all models produce unbiased estimates, i.e. the estimated interaction effect showed no substantial deviation from zero. Therefore these means are not reported here. Relative parameter bias for conditions with interaction effect are available as online supplementary material (Table 8).There is variation across models in parameter bias, especially for the smallest sample size. According to the criterion used by

Hoogland and Boomsma (1998) (relative bias less than 0.05 in absolute value) acceptable bias occurs in 67 of the 122 conditions with sample $n = 200$, and in 121 of the 122 conditions with $n = 926$. Non-normality of the data does not influence relative bias.

We also report the performance of standard error estimation for the 61 models as online supplementary material (Table 8). The mean estimated standard error is denoted by SE, while SD is the standard deviation of the estimated parameter $\hat{\gamma}_3$. Both SE and SD are calculated over all replications in each design cell. There is a tendency that SE is slightly underestimated, i.e. that the estimated standard error is less than the true value. Hoogland and Boomsma (1998) deemed standard error estimation to be acceptable if the relative bias in standard errors is below 0.1 in absolute value. According to this criterion, standard error estimation is acceptable in 89 of the 122 conditions with small sample size, and in all 122 conditions with $n = 926$. Similar to parameter bias, standard error estimation is not influenced by non-normality.

Note also that there is variation in efficiency among the models. That is, with some models the estimation of $\hat{\gamma}_3$ is obtained with more precision in terms of lower SD. Aggregating over sample sizes and normality conditions, the ALL configuration has the highest efficiency.

Coverage probabilities for 95% confidence intervals are presented in as online supplementary material (Table 9). These probabilities vary among configurations, particularly at the smallest sample size. Corresponding to the underestimation of standard errors, the coverage probabilities are generally too low compared to the nominal 95% level. We deem a coverage rate to be inadequate if it drops below 90% (see, e.g. Collins, Schafer, & Kam, 2001) . This occurs in only 33 of the 244 conditions. Note however, that at $n = 200$ the ALL configuration performs less well than most matched-pair configurations in terms of coverage.

We conclude that the PI approach generally performs well in terms of bias, efficiency and coverage. However, there is considerable variation among the various PI configurations.

Type I error. A central issue in interaction modeling is to determine whether an interaction effect exists. The hypotheses are $H_0 : \gamma_3 = 0$ against the alternative $H_1 : \gamma_3 \neq 0$. Rejection occurs when the z-value exceeds the critical value 1.96, that is, we set the significance level to $\alpha = .05$. Under the absence of an interaction effect, rejection of the null is a type I error, and should occur in 5% of the replications. Monte Carlo results concering type I errors and power are tabulated in online supplementary materials (Table 10). Across all conditions and configurations, the type I error rate was 4.58%, i.e. the PI approach in general tended to be conservative, rejecting a true null less often than the nominal rate. Distribution had an effect on type I error. For normal and non-normal data the type I error rates were 4.74% and 4.42%, respectively, aggregating over sample size and models. Aggregating over all models, sample size did not affect type I error under non-normal data. The impact of sample size in normal data was unexpected, with type I error rates aggregated over models of 4.97% and 4.52% for $n = 200$ and $n = 926$, respectively.

Next we discuss overall type I error rates for each model, i.e. aggregated over sample size and distribution. There was variation across models, especially in 3MATCH, with overall type I error rates ranging from 1.7% to 5.4%. Also within 4MATCH there was variation among configurations, with overall type I error rates ranging from 3.7% to 5.5%. The 3MATCH and 4MATCH configurations had aggregated type I error rates of 4.3% and 4.7%, respectively. The all-pairs configuration tended to have inflated type I error rates in small samples. The 3M18 configuration, on the other hand, is an instance of a matched-pair configuration with unacceptably low type I error rates.

Power. In the case where data was simulated with an interaction effect, the null $H_0 : \gamma_3 = 0$ is false. Provided type I error is acceptable, the higher power to detect interaction, the better. Distribution type had only a modest effect on power, overall power for non-normal and normal data was 70.5% and 71.9%, respectively. As expected, power increased markedly with sample size.

Next we consider overall power, aggregated over distribution and sample size, for

each configuration. The variation between configurations is significant, especially within 3MATCH. Variation decreased with increasing sample size. Overall power within 3MATCH and 4MATCH was 67.7% and 73.2%, respectively. The configuration with the highest overall power was the all-pair configuration.

Discussion

In study 2 we have seen considerable variation across PI configurations regarding convergence, bias, coverage, type I error and power. A general observation is that increasing the number of PIs leads to less variation among candidate configurations, as well as to superior performance in terms of convergence, type I error and power to detect interaction. Based on these observations we would recommend 4MATCH configuration over 3MATCH configurations. However, there is still a great deal of variation across 4MATCH configurations. For instance, in a typical sample the range of z-values obtained across 4MATCH configurations is quite large. The implication is that inference concerning the interaction effect could well be directly affected by the specific choice of a 4MATCH configuration.

A remedy for this variability is offered by the all-pairs configuration. The use of this configuration in interaction modeling has the great advantage of unambiguity. There is only one all-pairs configuration. A researcher choosing to use some other configuration must defend the specific choice of configuration over all other comparable configurations.

A further argument in favor of the all-pairs configuration is that it consistently performs well relative to the matched-pair configurations. It has the best convergence rates, and performs better than 50 of the 60 matched-pair configurations in terms of standard error bias. In terms of type I error and power it has acceptable performance, with one notable exception: Type I errors under the all-pairs configuration seems to be inflated under small sample size. The relatively poor performance at the small sample size might be explained by the fact that the all-pair configuration is more complex than matched-pair

configurations. Previous research has shown that when sample size is small relative to model complexity, standard errors may become attenuated (Chou, Bentler, & Satorra, 1991). The number of free parameters with the all-pairs configuration is $q = 87$. For 3MATCH and 4MATCH configurations the interaction model typically has $q = 33$, and $q = 37$, respectively. As more parameters are estimated under all-pairs, the required sample size is larger for this configuration. With $n = 200$ the number of observations per free parameter in all-pairs is critically low at $n : q = 2.3$. The corresponding ratio for 3MATCH configurations is $n : q = 6.1$.

Study 3: Monte Carlo Evaluation of All-pairs for Small Sample Sizes

Method

The previous studies imply that the PI configuration containing all pairs should be preferred under the PI approach. However, all-pairs starts to perform worse when sample size decreases. Whereas study 2 was limited to only two sample sizes, in study 3 we further study the performance of all-pairs under small sample sizes. In addition we compare all-pairs with the currently most popular non-PI method for modelling latent interactions, namely the LMS approach. Adhering to the post-hoc principle we include the following conditions.

- Interaction effect: none ($\gamma_3 = 0$) or real-world ($\gamma_3 = 0.15$).
- Distribution of data: normal (D1), real-world (D2), and severely non-normal (D3).
- Sample size: $n = 100, 200, 400$, and real-world $n = 926$.

The design yields 24 conditions, with 500 sample replications in each condition. Each random sample was generated as described in study 2. In addition to normal data (D1) and the distribution similar to the real-world case (D2), we added a more extreme condition of non-normality (D3), by setting the skewness and kurtosis of both PBC and ATT to 3 and 25, respectively. All other random constituents were kept normal. LMS estimation was done in MPLUS (Muthén & Muthén, 2010) through the use of the

MplusAutomation package (Hallquist & Wiley, 2013).

Results

Table 11 gives parameter bias (absolute for $\gamma_3 = 0$ and relative for $\gamma_3 = 0.15$), relative standard error bias, coverage and finally rejection rates for testing the significance of γ_3.

All-pairs provides unbiased estimates of γ_3 in all conditions. However, at $n = 100$ the bias in standard error estimation is unacceptable. With increasing sample size standard error bias is reduced, being acceptable in most conditions at $n = 200$. Standard error estimation is sensitive to degree of non-normality. Coverage rates for all-pairs is generally too low, especially for data with an interaction effect. Type I error rates (RR) are acceptable at sample sizes $n = 400, 926$, but are inflated at lower sample sizes. With increasing non-normality of data (D1 to D3) all-pairs performs worse, but overall we deem all-pairs to be quite robust to non-normality.

The LMS estimator has unbiased estimates of γ_3 in almost all conditions. A noticeable exception is when data have interaction and are highly non-normal (D3). In these conditions sample size does not seem to reduce the bias. LMS provides acceptable estimation of standard errors even at small sample sizes. Coverage is somewhat low in all conditions. Type I errors are generally inflated, even at large sample sizes.

Discussion

Study 3 was intended to map out a major weakness of the all-pairs configuration, namely the poor performance under small sample sizes. Although unbiased, all-pairs provides unacceptable bias in standard errors at $n = 100$, resulting in inflated Type I error rates and low coverage. These observations apply partially to $n = 200$, while for $n = 400$ all-pairs performed generally satisfactory. For $n = 400$ the $n : q$ ratio is 4.6, suggesting that $n : q$ ratios near 5 might be sufficient for all-pairs to perform satisfactory.

A second motivation for study 3 was to compare all-pairs with a non-PI approach. In terms of standard error estimation LMS outperformed all-pairs. However, in the most

non-normal situation (D3) parameter bias was unacceptable. LMS being a ML estimator based on the normality assumption, it is in fact inconsistent under non-normality. Generally, LMS had better coverage than all-pairs. However, at all but the smallest sample size all-pairs had slightly better Type I error control than LMS. Overall we may conclude that LMS performs as well or better than all-pairs in many conditions, but that all-pairs is preferable under the non-normal data condition D3. Note also that for the present real-world case, with $n = 926$, D2 and $\gamma_3 = 0.15$, all-pairs is preferable to LMS.

General Discussion

The application of post hoc modeling has gained interest in recent years (Bandalos, 2006; Hancock, 2006; MacCallum, 2003). The present study applied post hoc analysis in two ways to study how the choice of product indicators to operationalize the latent interaction variable affects estimation and interaction inference.

In the first study, variation both in convergence and in obtained interaction effect across models are contingent upon a single real-world dataset. All possible configurations were assessed with respect to variation of the interaction parameter value and its significance. To complement the findings of the first post hoc study, a post hoc Monte Carlo evaluation based on the same real-world sample was conducted. In the second study variation across all sixty matched-pair configurations in a typical sample was assessed by calculating the mean range over all replications in each of the eight design cells. Study 2 in addition provided for each of the 60 matched-pair configurations, and for the all-pairs configuration, information about convergence rates, standard error bias, Type 1 error, and power to detect the interaction under different design conditions. Common for both approaches in study 2 is the focus on how the 61 configurations vary with respect to assessing the interaction parameter γ_3. It may be noted that the Monte Carlo study is restricted in scope due to its post hoc nature, while on the other hand it provides a sampling perspective that allows calculation of standard error bias, Type 1 error and power

in addition to convergence rates.

Given the present three and four indicators for the two first-order factors, respectively, there are 4083 different configurations or operational definitions of the latent interaction variable if we exclude configurations with only one PI. Configurations differ with respect to the number of PIs they contain and to which extent they balance the information contained in the first-order indicators. That is, a configuration is unbalanced in the sense that it reuses the same information several times at the expense of excluding some information. For example the configuration $p_1a_1, p_1a_2, p_1a_3, p_1a_4$ puts a disproportionate weight on p_1 at the expense of p_2 and p_3. It is unlikely that an applied researcher would favor unbalanced configurations over balanced configurations. In balanced configurations the available indicators of the two first-order factors are more evenly distributed throughout the available PIs. From a content validity perspective (Ping, 1998; Kane, 2001, 2006) a balanced configurations would be preferred in favor of an unbalanced configuration. Prior analyses, carried out by the present authors, suggested that balanced configurations in general outperformed unbalanced configurations with respect to convergence and type I error control. The present study focused on balanced configurations of the matched-pair type, which are well-known and recommended in the literature (Marsh et al., 2004; Marsh, Wen, Nagengast, & Hau, 2012).

The results from the first post hoc analysis suggested that the likelihood of obtaining a proper solution increased as the number of PIs increased. Substantial amount of variation among configurations with respect to the estimated interaction parameter, as well as to its z-value, were demonstrated. Even though the 4MATCH configurations displayed less variation than the 3MATCH configurations, considerable variation was detected across the 4MATCH configurations. Similar findings were obtained in the Monte Carlo study. When aggregating over the eight design conditions, convergence rates were lower for 3MATCH than for 4MATCH. The all-pairs configuration displayed an overall convergence rate that was higher than all of the 3- and 4-MATCH configurations. These results do not

support the concern raised by Marsh et al. (2004) that nonconvergence problems may occur by increasing the number of PIs. Thus a consistent finding obtained in the present study is that generally the PI approach performed better as the number of PIs increased.

Considerable variation across the present PI configurations were generally displayed regarding convergence, parameter estimates, standard error bias, type 1 error and power. Even though variation across the PI configuration decreased as number of PIs increased, i.e. with less variation across 4MATCH than across 3MATCH, we deem even the former variation to be substantial. An applied researcher choosing among 4MATCH configurations would face unacceptable levels of ambiguity. Because the present results cannot offer sufficient credibility to recommending even the 4MATCH configurations, our choice of recommendation would therefore be to use the all-pairs configuration. With this type of configuration there is no ambiguity, since the all-pairs configuration is unique. Any other choice of configuration will eventually involve the researcher choosing, more or less arbitrarily, a set of PIs. Our recommendation is strengthened by the fact that the all-pairs configuration demonstrated the best convergence rate and provided better standard error bias than 50 of the 60 mathed-pairs configurations. It also performed well in terms of type I error control and power.

However, the recommendation of all-pairs cannot be given without reservations. It is helpful to frame the following discussion in terms of the observations-to-parameters ratio $n : q$, where n and q denote sample size and the number of parameters in the model, respectively. Low values of $n : q$, say below 5, indicate that the sample may be small relative to the model complexity (see, e.g., Nevitt & Hancock, 2004). The unhealthy combination of small samples and complex models with relatively large number of parameters was observed in studies 2 and 3, where the all-pairs had unacceptable performance in terms of standard error bias and type I error at $n = 100, 200$, but performed well at $n = 400$. With 87 free parameters the $n : q$ ratios for $n = 100, 200$ are 1.15 and 2.3, respectively. It follows that use of all-pairs in the present real-world setting,

with 87 free parameters, is not recommended at the smallest sample sizes $n = 100, 200$. However, we expect that all-pairs will provide acceptable type I error control with a higher observations-to-parameters ratio of, say, $n : q = 5$ or higher. In fact, in study 3 we found that performance was acceptable at $n = 400$, which has $n : q = 4.6$.

Even with a large sample $n : q$ may be low, if the model has many free parameters. For the PI approach this may occur if the latent constructs each have many indicators. To illustrate the growing complexity of the all-pairs configuration as the number of first-order indicators increases, remark that with eight indicators for each first-order factor, there are 448 correlated uniquenesses. In the present study the first-order factors had three or four indicators, resulting in a rather uncomplicated all-pairs configuration demanding 30 correlated uniquenesses due to the sharing of common first-order indicators. It may be argued that for many psychological constructs three or four indicators represent a rather narrow selection of indicators to adequately represent the construct in question. Small sets of indicators may satisfy internal consistency reliability, but if they are considered too narrow for the construct in mind, validity will suffer (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Nunnally & Bernstein, 1994). However, if the sets of first-order indicators suffer from weak content validity, it follows that the operational definition of the latent interaction will also suffer (Ping, 1998). A remedy for the increasing complexity of PI configurations with increasing number of first-order indicators may be sought through application of parcels to reduce the number of indicators. However, making parcels introduces the same kind of arbitrariness that occur in the PI approach. In general, parceling continues to be a controversial issue (Bandalos, 2002; Kim & Hagtvet, 2003; Sterba & MacCallum, 2010; Sterba, 2011; Marsh, Lüdtke, Nagengast, Morin, & Von Davier, 2013; Little, Rhemtulla, Gibson, & Schoemann, 2013). The challenge of creating parcels for estimating latent variable interaction that satisfy validity concerns as well as render possible estimation has been rarely addressed (one exception is Jackman, Leite, and Cochrane (2011)). This issue should be addressed in future research.

Conclusion

Currently no clear consensus exists concerning the number and type of PIs to apply in the PI approach to latent variable interaction modeling. In this study we examined the effect the choice of PIs has on the estimated interaction effect in a real-world substantive research effort with a single dataset. The present findings are limited to our real-world empirical case. However, this case may be as representative of a practical research situation as the scenario defined in pure Monte Carlo studies with artificially generated samples.

Overall, we conclude that within the PI approach the all-pairs configuration has the best statistical properties. A viable option that does not rely on PIs is the LMS approach, which was found to perform as well or better than the all-pairs configuration in optimal conditions. However, for severely non-normal data all-pairs is preferable.

A main advantage of the all-pairs configuration is its unambiguity. That is, there is only one all-pairs configuration in any modeling context. This leaves no choice for researchers to search among different plausible configurations for a model that supports their conjecture. To claim valid interpretation of estimated latent interactions requires both sound statistical methodology as well as reasonable conceptual considerations. Within the PI approach the present study suggests that the all-pairs configuration will best serve this purpose.

References

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling*, *9*(1), 78–102.

Bandalos, D. L. (2006). The use of monte carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 385–426). Greenwich, CT: Information Age. Greenwich, CT.

Chou, C.-P., Bentler, P. M., & Satorra, A. (1991). Scaled test statistics and robust standard errors for non-normal data in covariance structure analysis: a monte carlo study. *British Journal of Mathematical and Statistical Psychology*, *44*(2), 347–357.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods*, *6*(4), 330.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Friedrich, R. J. (1982). In defense of multiplicative terms in multiple regression equations. *American Journal of Political Science*, 797–833.

Hallquist, M., & Wiley, J. (2013). Mplusautomation: Automating mplus model estimation and interpretation [Computer software manual].

Hancock, G. R. (2006). Power analysis in covariance structure modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 69–115). Greenwich, CT: Information Age Publishing.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling an overview and a meta-analysis. *Sociological Methods & Research*, *26*(3), 329–367.

Hukkelberg, S. S., Hagtvet, K. A., & Kovac, V. B. (2013). Latent interaction effects in the theory of planned behaviour applied to quitting smoking. *British journal of health*

psychology.

Jackman, M. G.-A., Leite, W. L., & Cochrane, D. J. (2011). Estimating Latent Variable Interactions with the Unconstrained Approach: A Comparison of Methods to Form Product Indicators for Large, Unequal Numbers of Items. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(2), 274–288.

Jöreskog, K., & Yang, F. (1996). Nonlinear structural equation models: The Kenny-Judd model with interaction effects. In G. Marcoulides & R. Schumacher (Eds.), *Advanced structural equation modeling: Issues and techniques* (p. 57-88). Mahwah, NJ: Lawrence Erlbaum Associates.

Kane, M. T. (2001). Current concerns in validity theory. *Journal of educational Measurement*, *38*(4), 319–342.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education/Praeger.

Kenny, D., & Judd, C. (1984). Estimating the nonlinear and interactive effects of latent variables. *Psychological Bulletin*, *96*(1), 201–210.

Kim, S., & Hagtvet, K. A. (2003). The impact of misspecified item parceling on representing latent variables in covariance structure modeling: A simulation study. *Structural Equation Modeling*, *10*(1), 101–127.

Klein, A., & Moosbrugger, H. (2000). Maximum likelihood estimation of latent interaction effects with the lms method. *Psychometrika*, *65*(4), 457–474.

Klein, A., & Muthén, B. (2007). Quasi-Maximum Likelihood Estimation of Structural Equation Models With Multiple Interaction and Quadratic Effects. *Multivariate Behavioral Research*, *42*(4), 647–673.

Lin, G.-C., Wen, Z., Marsh, H., & Lin, H.-S. (2010, July). Structural Equation Models of Latent Interactions: Clarification of Orthogonalizing and Double-Mean-Centering Strategies. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(3), 374–391.

Little, T., Bovaird, J., & Widaman, K. (2006). On the Merits of Orthogonalizing Powered and Product Terms: Implications for Modeling Interactions Among Latent Variables. *Structural Equation Modeling: A Multidisciplinary Journal*, *13*(4), 497–519.

Little, T., Rhemtulla, M., Gibson, K., & Schoemann, A. (2013). Why the items versus parcels controversy needn't be one. *Psychological Methods*, *18*(3).

Luszczynska, A., & Schwarzer, R. (2003). Planning and self-efficacy in the adoption and maintenance of breast self-examination: A longitudinal study on self-regulatory cognitions. *Psychology and Health*, *18*, 93–108.

MacCallum, R. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research*, *38*(1), 113–139.

Marsh, H. W., Lüdtke, O., Nagengast, B., Morin, A. J., & Von Davier, M. (2013). Why item parcels are (almost) never appropriate: Two wrongs do not make a right—camouflaging misspecification with item parcels in cfa models. *Psychological methods*, *18*(3), 257.

Marsh, H. W., Wen, Z., & Hau, K. T. (2004). Structural Equation Models of Latent Interactions: Evaluation of Alternative Estimation Strategies and Indicator Construction. *Psychological Methods*, *9*(3), 275–300.

Marsh, H. W., Wen, Z., & Hau, K. T. (2007). Unconstrained Structural Equation Models of Latent Interactions: Contrasting Residual- and Mean-Centered Approaches. *Structural Equation Modeling: A Multidisciplinary Journal*, 1–12.

Marsh, H. W., Wen, Z., Nagengast, B., & Hau, K. T. (2012). Structural Equation Models of Latent Interaction. In R. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 436–458). Springer.

Mooijaart, A., & Bentler, P. (2010, July). An Alternative Approach for Nonlinear Latent Variable Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(3), 357–373.

Muthén, L., & Muthén, B. (2010). *Mplus software (version 6.1).* Los Angeles, CA: Muthén

Muthén.

Nevitt, J., & Hancock, G. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research*, *39*(3), 439–478.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Ping, R. (1998). Eqs and lisrel examples using survey data. In R. E. Schumacker & G. A. Marcoulides (Eds.), *Interaction and nonlinear effects in structural equation modeling* (pp. 63–100). Mahwah, NJ: Lawrence Erlbaum.

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*(3), 354–373.

Rosseel, Y. (2012). lavaan: Latent variable analysis [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=lavaan` (R package version 0.4-12)

Saris, W. E., Batista-Foguet, J. M., & Coenders, G. (2007, February). Selection of Indicators for the Interaction Term in Structural Equation Models with Interaction. *Quality & Quantity*, *41*(1), 55–72.

Satorra, A., & Bentler, P. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. V. Eye & C. Clogg (Eds.), *Latent variable analysis: applications for developmental research* (p. 399-419). Newbury Park, CA: Sage.

Steinmetz, H., Davidov, E., & Schmidt, P. (2011). Three approaches to estimate latent interaction effects: Intention and perceived behavioral control in the theory of planned behavior. *Methodological Innovations Online*, *6*(1), 95–110.

Sterba, S. K. (2011). Implications of parcel-allocation variability for comparing fit of item-solutions and parcel-solutions. *Structural Equation Modeling: A*

Multidisciplinary Journal, *18*(4), 554–577.

Sterba, S. K., & MacCallum, R. C. (2010). Variability in parameter estimates and model fit across repeated allocations of items to parcels. *Multivariate Behavioral Research*, *45*(2), 322–358.

Wall, M. M., & Amemiya, Y. (2001). Generalized appended product indicator procedure for nonlinear structural equation analysis. *Journal of Educational and Behavioral Statistics*, *26*, 1–29.

Wen, Z., Marsh, H. W., & Hau, K.-T. (2010, January). Structural Equation Models of Latent Interactions: An Appropriate Standardized Solution and Its Scale-Free Properties. *Structural Equation Modeling: A Multidisciplinary Journal*, *17*(1), 1–22.

Yang-Wallentin, F. (1998). Modeling interaction and nonlinear effects: A step-by-step lisrel example. In R. E. Schumacker & G. A. Marcoulides (Eds.), *Interaction and nonlinear effects in structural equation modeling* (pp. 1–26). Mahwah, NJ: Lawrence Erlbaum.

Yang-Wallentin, F., Schmidt, P., Davidov, E., & Bamberg, S. (2004). Is there any interaction effect between intention and perceived behavioral control. *Methods of Psychological Research Online*, *8*(2), 127–157.

Table 1

Correlation Matrix, Mean, Standard Deviation, Skewness and Kurtosis of Indicator Variables

	p_1	p_2	p_3	a_1	a_2	a_3	a_4	g_1	g_2	g_3
p_1	1.00									
p_2	.83	1.00								
p_3	.69	.66	1.00							
a_1	.12	.14	-.09	1.00						
a_2	.05	.09	-.12	.61	1.00					
a_3	.29	.30	.16	.39	.25	1.00				
a_4	.17	.18	-.04	.60	.43	.56	1.00			
g_1	.34	.33	.15	.31	.21	.26	.31	1.00		
g_2	.33	.33	.17	.32	.22	.23	.30	.77	1.00	
g_3	.31	.28	.19	.26	.17	.23	.28	.62	.71	1.00
Mean	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
SD	1.8	1.8	1.7	1.5	1.1	2.2	2.0	2.0	2.0	1.7
Skewness	-.1	-.1	.9	-1.7	-2.9	-.1	-.8	.8	.8	1.0
Kurtosis	-.9	-1.0	-.2	2.1	8.9	-1.4	-.8	-.7	-.6	.2

Note. x_i, x_j and g_k are indicators for Perceived Behavioral Control, Attitude and Goal Commitment, respectively.

Table 2

Proper and Improper Solutions for Different Number of PIs

Number of PI's	2	3	4	5	6	7	8	9	10	11	12
Proper	23	126	345	633	823	753	486	219	66	12	1
Improper	43	94	150	159	101	39	9	1	0	0	0
Percentage proper	35	57	70	80	89	95	98	99.5	100	100	100

Table 3

Incidence matrix for configurations included in study 2.

Model	P11	P12	P13	P14	P21	P22	P23	P24	P31	P32	P33	P34
$3M_{1}$	★					★					★	
$3M_{2}$	★					★						★
$3M_{3}$	★						★			★		
$3M_{4}$	★						★					★
$3M_{5}$	★							★		★		
$3M_{6}$	★							★			★	
$3M_{7}$		★			★						★	
$3M_{8}$		★			★							★
$3M_{9}$		★					★		★			
$3M_{10}$		★					★					★
$3M_{11}$		★						★	★			
$3M_{12}$		★						★			★	
$3M_{13}$			★		★					★		
$3M_{14}$			★		★							★
$3M_{15}$			★			★			★			
$3M_{16}$			★			★						★
$3M_{17}$			★					★	★			
$3M_{18}$			★					★		★		
$3M_{19}$				★	★					★		
$3M_{20}$				★	★						★	
$3M_{21}$				★		★			★			
$3M_{22}$				★		★					★	
$3M_{23}$				★			★		★			
$3M_{24}$				★			★			★		
$4M_{1}$	★	★					★					★
$4M_{2}$	★	★						★			★	
$4M_{3}$	★		★			★						★
$4M_{4}$	★		★					★		★		
$4M_{5}$	★			★		★					★	
$4M_{6}$	★			★			★			★		
$4M_{7}$	★					★	★					★
$4M_{8}$	★					★		★			★	
$4M_{9}$	★					★					★	★
$4M_{10}$	★						★	★		★		
$4M_{11}$	★						★			★		★
$4M_{12}$	★							★		★	★	
$4M_{13}$		★	★		★							★
$4M_{14}$		★	★					★	★			
$4M_{15}$		★		★	★						★	
$4M_{16}$		★		★			★		★			
$4M_{17}$		★			★		★					★
$4M_{18}$		★			★			★			★	
$4M_{19}$		★			★						★	★
$4M_{20}$		★					★	★	★			
$4M_{21}$		★					★		★			★
$4M_{22}$		★						★	★		★	
$4M_{23}$			★	★	★					★		
$4M_{24}$			★	★		★			★			
$4M_{25}$			★		★	★						★
$4M_{26}$			★		★			★		★		
$4M_{27}$			★		★					★		★
$4M_{28}$			★			★		★	★			
$4M_{29}$			★			★			★			★
$4M_{30}$			★					★	★	★		
$4M_{31}$				★	★	★					★	
$4M_{32}$				★	★		★			★		
$4M_{33}$				★	★					★	★	
$4M_{34}$				★		★	★		★			
$4M_{35}$				★		★			★		★	
$4M_{36}$				★			★		★	★		
ALL	★	★	★	★	★	★	★	★	★	★	★	★

Table 4

Non-normality condition. Skewness and kurtosis in indicators, calculated from a simulated sample with $n = 10^6$.

	p_1	p_2	p_3	a_1	a_2	a_3	a_4	g_1	g_2	g_3
Skewness	-.4	-.4	-.2	-.7	-1.4	-.1	-.3	.0	.0	.0
Kurtosis	-.8	-.6	-.3	4.8	4.0	.2	1.2	.1	.1	.0

Table 5

Range, Max and Min of $\tilde{\gamma}_3$ across Configurations in a Typical Sample

		$\gamma_3 = 0$				$\gamma_3 = .15$			
		Non-normal		Normal		Non-normal		Normal	
Config	n	200	926	200	926	200	926	200	926
3MATCH	Range	.36	.09	.27	.08	.53	.14	.43	.13
	Max	. 17	.04	.14	.04	.52	.25	.44	.24
	Min	-.20	-.04	-.13	-.04	-.01	.11	.01	.11
4MATCH	Range	.18	.06	.17	.06	.28	.09	.27	.09
	Max	.09	.03	.09	.03	.34	.22	.32	.22
	Min	-.09	-.03	-.08	-.03	.06	.13	.05	.12

Note. The max and min of $\tilde{\gamma}_3$, together with range=max-min, was calculated across configurations in each replicated sample. Tabulated values are means over all replicated samples in the cell.

Table 6

Range, Max and Min of z across Configurations in a Typical Sample

		$\gamma_3 = 0$				$\gamma_3 = .15$			
		Non-normal		Normal		Non-normal		Normal	
Config	n	200	926	200	926	200	926	200	926
3MATCH	Range	1.61	1.58	1.57	1.51	2.07	3.00	2.08	2.87
	Max	.82	.78	.80	.76	2.54	4.78	2.60	4.83
	Min	-.80	-.80	-.76	-.76	.47	1.77	.52	1.96
4MATCH	Range	1.54	1.31	1.50	1.29	1.80	2.23	1.84	2.26
	Max	.78	.66	.76	.66	2.59	4.81	2.64	4.86
	Min	-.76	-.65	-.74	-.64	.79	2.58	.80	2.61

Note. The max and min of z, together with range=max-min, was calculated across configurations in each replicated sample. Tabulated values are means over all replicated samples in the cell.

Table 7

Percentage of Converged Solutions. γ_3*= interaction effect.*

	$\gamma_3 = 0$				$\gamma_3 = 0.15$				
	Non-normal		Normal		Non-normal		Normal		
n	200	926	200	926	200	926	200	926	Overall
$3M_1$	95.5	99.9	95.7	100.0	96.5	100.0	97.8	100	98.2
$3M_2$	98.4	100	99.4	100	99.2	100	99.7	100	99.6
$3M_3$	95	100	96.6	100	97	100	98.1	100	98.3
$3M_4$	71.9	83	66.7	69.5	85.5	98.9	84.6	99.7	82.5
$3M_5$	98.6	100	99.5	100	99.1	100	99.7	100	99.6
$3M_6$	75	86.2	71.2	72.5	88.8	99.9	88	99.8	85.2
$3M_7$	95	100	95.6	100	95.5	100	96.8	100	97.9
$3M_8$	99	100	99.5	100	98.9	100	99.7	100	99.6
$3M_9$	96.6	100	97.6	100	97.5	100	98.5	100	98.8
$3M_{10}$	70.6	78	64	64.1	81	98.6	80	99	79.4
$3M_{11}$	98.8	100	99.8	100	99.1	100	99.9	100	99.7
$3M_{12}$	69.7	80.1	65.3	66.9	81.5	99.3	81	98.7	80.3
$3M_{13}$	95.3	100	97.2	100	96.3	100	97.7	100	98.3
$3M_{14}$	72.1	80	69.4	68.8	86.8	99.3	85.2	99.4	82.6
$3M_{15}$	97.3	100	98	100	98	100	98.2	100	98.9
$3M_{16}$	66.3	75.1	64.6	63.9	81.2	98.4	80.4	98.8	78.6
$3M_{17}$	69.3	79.5	68.1	68.6	84.9	98.4	84.7	99.7	81.7
$3M_{18}$	65	75.2	62.1	63.5	76.3	96.6	77.8	97.6	76.8
$3M_{19}$	98.8	100	99.7	100	99	100	99.8	100	99.7
$3M_{20}$	77.2	85.9	71	70.6	89.6	99.9	88.6	99.8	85.3
$3M_{21}$	99	100	99.7	100	98.9	100	99.8	100	99.7
$3M_{22}$	70.3	79.5	66	66.7	82.7	99.3	82.5	99.1	80.8
$3M_{23}$	71.1	81.4	69.2	69.2	85.8	98.7	86.2	99.6	82.7
$3M_{24}$	66	76	63.5	63	75.6	97.3	78	98	77.2
$4M_1$	72.7	83.8	67.1	72.1	87	98.9	85.4	99.7	83.3
$4M_2$	75.4	86.3	70.7	73.1	88.8	99.8	89.3	99.8	85.4
$4M_3$	98.5	100	99.4	100	99.2	100	99.8	100	99.6
$4M_4$	98.6	100	99.6	100	98.9	100	99.7	100	99.6
$4M_5$	95.1	99.9	95.4	100	96.3	100	97.7	100	98.0
$4M_6$	93.3	100	95.8	100	95.7	100	97.8	100	97.8
$4M_7$	98.4	100	99.2	100	99.2	100	99.6	100	99.5
$4M_8$	95.1	99.9	95.6	100	96.9	100	97.8	100	98.2
$4M_9$	98.3	100	98.9	100	98.7	100	99.2	100	99.4
$4M_{10}$	98.4	100	99.2	100	98.8	100	99.6	100	99.5
$4M_{11}$	94.1	100	96.4	100	95.7	100	97.7	100	98.0
$4M_{12}$	98.4	100	99.4	100	99.2	100	99.8	100	99.6
$4M_{13}$	99	100	99.3	100	98.9	100	99.8	100	99.6
$4M_{14}$	98.9	100	99.8	100	99.4	100	99.9	100	99.8
$4M_{15}$	95.9	100	96.7	99.9	98	100	98.7	100	98.7
$4M_{16}$	95.4	100	97.3	100	97	100	98.8	100	98.7
$4M_{17}$	98.5	100	99.5	100	98.8	100	99.7	100	99.6
$4M_{18}$	94.7	100	95.8	100	96.7	100	97.7	100	98.1
$4M_{19}$	98.6	100	99.2	100	98.5	100	99.5	100	99.5
$4M_{20}$	98.6	100	99.7	100	99.2	100	100	100	99.7
$4M_{21}$	95.6	100	96.9	100	97	100	98.4	100	98.5
$4M_{22}$	98.8	100	99.8	100	99	100	99.9	100	99.7
$4M_{23}$	98.7	100	99.5	100	99.1	100	99.5	100	99.6
$4M_{24}$	98.7	100	99.4	100	98.9	100	99.6	100	99.6
$4M_{25}$	72.8	80.4	68.5	68.2	87.3	99.3	86.3	99.6	82.8
$4M_{26}$	94.9	99.9	96.2	100	96.7	100	97.3	100	98.1
$4M_{27}$	94.6	100	96.7	100	97.1	100	97.9	100	98.3
$4M_{28}$	97.2	100	97.5	100	98.2	100	98.5	100	98.9
$4M_{29}$	96.7	100	97.1	100	97.9	100	98.2	100	98.7
$4M_{30}$	70.5	79.5	68.1	67.3	85.2	98.3	86.3	99.6	81.8
$4M_{31}$	78	85.8	71.8	70.7	90.1	99.9	88.9	99.8	85.6
$4M_{32}$	98.5	100	99.7	100	98.6	100	99.7	100	99.6
$4M_{33}$	98.5	100	99.8	100	99	100	100	100	99.7
$4M_{34}$	99.3	100	99.7	100	99.1	100	100	100	99.8
$4M_{35}$	98.9	100	99.7	100	99.1	100	99.8	100	99.7
$4M_{36}$	71.9	81.5	69	68.8	86.3	99.1	86.7	99.6	99.4
ALL	99.4	100	99.9	100	99.9	100	100	100.0	99.9

Table 8

Estimation of γ_3 when $\gamma_3 \neq 0$: relative bias and estimated and empirical standard errors.

RB=relative bias. SE= estimated standard error. SD= empirical standard error.

	Non-normal						Normal					
n	3c200			926			200			926		
	RB	SE	SD	RB	SE	SD	RB	SE	SD	RB	SE	SD
$3M_1$	0	.087	.093	-.01	.039	.038	0	.078	.088	0	.036	.035
$3M_2$	-.02	.077	.081	0	.035	.034	0	.072	.077	0	.033	.032
$3M_3$	.01	.093	.096	-.01	.04	.039	0	.083	.091	0	.038	.036
$3M_4$	.1	.108	.109	-.01	.051	.051	.14	.107	.106	0	.051	.052
$3M_5$	-.02	.076	.08	-.01	.035	.033	0	.073	.078	0	.033	.032
$3M_6$	.07	.096	.1	-.03	.046	.046	.08	.097	.102	-.01	.046	.046
$3M_7$	.02	.136	.15	.01	.062	.059	.02	.129	.14	.01	.06	.057
$3M_8$	.04	.138	.146	.01	.061	.058	.04	.131	.138	.01	.059	.056
$3M_9$	.04	.161	.175	.02	.07	.069	.03	.148	.158	.02	.067	.066
$3M_{10}$	.14	.253	.308	.02	.097	.104	.15	.222	.253	.02	.095	.103
$3M_{11}$	.06	.155	.166	.02	.066	.065	.06	.145	.15	.02	.063	.062
$3M_{12}$	.08	.211	.29	0	.087	.09	.15	.195	.238	.01	.086	.09
$3M_{13}$	.07	.129	.14	.01	.053	.051	.09	.136	.155	.01	.055	.052
$3M_{14}$	.09	.133	.142	.01	.055	.054	.12	.148	.181	.01	.057	.055
$3M_{15}$	.11	.152	.182	.03	.06	.059	.09	.151	.158	.02	.06	.058
$3M_{16}$	.41	.383	.496	.02	.088	.089	.18	.261	.239	.01	.079	.08
$3M_{17}$	.15	.184	.199	.03	.066	.067	.09	.173	.178	.01	.067	.068
$3M_{18}$	.8	.802	1.299	.05	.107	.107	.42	.446	.669	.03	.093	.093
$3M_{19}$	.02	.075	.079	.01	.034	.032	.03	.075	.078	.01	.034	.031
$3M_{20}$	.03	.077	.083	.01	.035	.034	.02	.075	.084	0	.035	.034
$3M_{21}$	.03	.084	.087	.01	.037	.036	.05	.081	.083	.01	.036	.035
$3M_{22}$	-.97	.139	3.377	0	.055	.055	.09	.112	.124	0	.048	.049
$3M_{23}$	.05	.092	.097	.02	.043	.044	.04	.09	.1	0	.042	.044
$3M_{24}$	.29	.234	.258	.02	.066	.066	.22	.155	.204	.01	.057	.057
$4M_1$	.07	.108	.11	-.01	.051	.051	.12	.107	.108	0	.051	.053
$4M_2$	.07	.097	.099	-.02	.046	.046	.07	.097	.103	-.01	.046	.046
$4M_3$	-.03	.076	.081	0	.035	.034	-.01	.072	.078	0	.033	.032
$4M_4$	-.03	.075	.081	-.01	.035	.033	-.02	.072	.078	-.01	.033	.032
$4M_5$	0	.086	.092	-.01	.039	.038	0	.078	.087	-.01	.036	.035
$4M_6$	0	.09	.094	-.01	.04	.039	0	.082	.089	0	.038	.036
$4M_7$	-.03	.075	.08	0	.035	.034	-.01	.071	.077	0	.033	.032
$4M_8$	-.03	.078	.085	-.01	.036	.035	0	.074	.083	-.01	.035	.033
$4M_9$	-.01	.076	.081	0	.035	.034	0	.072	.077	0	.033	.032
$4M_{10}$	-.02	.076	.079	-.01	.035	.033	-.01	.072	.078	0	.033	.032
$4M_{11}$	0	.086	.091	-.01	.039	.038	.01	.08	.086	0	.037	.036
$4M_{12}$	-.01	.076	.081	-.01	.035	.033	0	.072	.078	0	.033	.032
$4M_{13}$	.04	.136	.146	.01	.061	.058	.04	.13	.138	.01	.059	.056
$4M_{14}$	.05	.154	.166	.02	.066	.065	.06	.143	.151	.02	.063	.063
$4M_{15}$	.01	.132	.147	.01	.061	.058	.01	.128	.14	.01	.059	.057
$4M_{16}$	.02	.154	.165	.02	.068	.068	.02	.144	.155	.01	.065	.065
$4M_{17}$	.03	.135	.147	.01	.06	.058	.03	.129	.137	.01	.059	.056
$4M_{18}$	.02	.133	.147	.01	.061	.058	.01	.127	.14	.01	.059	.057
$4M_{19}$	.04	.137	.145	.01	.061	.058	.04	.13	.137	.01	.059	.056
$4M_{20}$	.04	.153	.162	.02	.066	.065	.05	.143	.146	.02	.063	.062
$4M_{21}$	.05	.159	.178	.02	.069	.069	.03	.146	.157	.01	.066	.066
$4M_{22}$	.05	.152	.17	.01	.066	.065	.05	.142	.149	.02	.063	.062
$4M_{23}$	.08	.133	.136	.01	.053	.05	.11	.142	.155	.02	.054	.052
$4M_{24}$	.12	.153	.168	.02	.058	.057	.12	.153	.153	.02	.058	.057
$4M_{25}$	.08	.13	.144	.01	.054	.053	.08	.138	.164	.01	.056	.055
$4M_{26}$	.09	.135	.168	.01	.053	.051	.1	.133	.166	.01	.055	.052
$4M_{27}$	.1	.136	.164	.01	.053	.051	.1	.14	.154	.01	.055	.052
$4M_{28}$	.07	.145	.169	.02	.058	.058	.11	.147	.171	.02	.059	.057
$4M_{29}$	.12	.149	.178	.03	.059	.059	.1	.147	.16	.02	.06	.058
$4M_{30}$	.2	.218	.255	.02	.065	.066	.12	.172	.209	.01	.066	.067
$4M_{31}$	.02	.075	.083	0	.035	.034	0	.073	.084	0	.035	.034
$4M_{32}$	.01	.074	.08	.01	.034	.032	.02	.074	.078	.01	.034	.031
$4M_{33}$	.03	.075	.08	.01	.034	.032	.04	.074	.078	.01	.034	.031
$4M_{34}$	.04	.084	.088	.01	.037	.036	.05	.081	.083	.01	.036	.035
$4M_{35}$	.03	.083	.088	.01	.037	.036	.04	.081	.084	.01	.036	.035
$4M_{36}$	.04	.092	.097	.01	.042	.044	.02	.087	.099	0	.041	.043
ALL	-.03	.07	.079	-.01	.033	.032	-.01	.068	.076	0	.032	.031

Table 9

Coverage rates of confidence intervals for γ_3. Confidence level 95 %.

	$\gamma_3 = 0$				$\gamma_3 = .15$			
	Non-normal		Normal		Non-normal		Normal	
n	200	926	200	926	200	926	200	926
$3M_1$	.956	.964	.955	.962	.892	.932	.885	.942
$3M_2$	.957	.954	.947	.952	.905	.941	.904	.944
$3M_3$	.966	.962	.96	.96	.897	.932	.885	.933
$3M_4$	.965	.967	.946	.938	.939	.936	.936	.934
$3M_5$	.958	.951	.94	.949	.895	.953	.893	.95
$3M_6$	.951	.976	.941	.948	.919	.941	.919	.94
$3M_7$	.948	.956	.937	.956	.91	.957	.906	.952
$3M_8$	.945	.958	.94	.956	.928	.96	.933	.958
$3M_9$	.944	.946	.942	.951	.921	.937	.926	.939
$3M_{10}$	.928	.955	.941	.938	.9	.928	.922	.925
$3M_{11}$	.949	.95	.939	.95	.936	.946	.941	.948
$3M_{12}$	.93	.949	.93	.952	.901	.941	.898	.929
$3M_{13}$	.949	.955	.949	.962	.937	.946	.925	.956
$3M_{14}$	.969	.955	.963	.953	.918	.951	.913	.947
$3M_{15}$	.95	.946	.957	.954	.927	.948	.93	.951
$3M_{16}$	.97	.983	.969	.964	.904	.946	.891	.935
$3M_{17}$	.962	.95	.963	.946	.928	.947	.897	.94
$3M_{18}$	.963	.98	.974	.975	.917	.947	.905	.944
$3M_{19}$	.947	.956	.94	.956	.93	.952	.94	.954
$3M_{20}$	.957	.946	.949	.949	.917	.949	.904	.943
$3M_{21}$	.942	.949	.955	.954	.933	.949	.94	.956
$3M_{22}$	.969	.967	.968	.957	.906	.945	.904	.932
$3M_{23}$	.941	.946	.964	.945	.924	.934	.905	.934
$3M_{24}$	.967	.966	.962	.957	.927	.937	.927	.935
$4M_1$	.957	.967	.937	.928	.93	.937	.932	.932
$4M_2$	.96	.964	.934	.936	.92	.945	.917	.941
$4M_3$	.956	.953	.947	.949	.899	.938	.9	.944
$4M_4$	.955	.948	.933	.946	.893	.948	.879	.95
$4M_5$	.955	.962	.954	.96	.892	.931	.887	.946
$4M_6$	.967	.962	.959	.957	.904	.928	.883	.933
$4M_7$	.957	.954	.948	.952	.904	.94	.902	.948
$4M_8$	.958	.96	.942	.96	.885	.941	.888	.944
$4M_9$	.956	.954	.946	.951	.9	.941	.896	.945
$4M_{10}$	.955	.952	.942	.948	.903	.953	.887	.948
$4M_{11}$	.956	.964	.952	.957	.91	.931	.889	.938
$4M_{12}$	.958	.952	.941	.948	.896	.952	.891	.949
$4M_{13}$	.947	.951	.94	.952	.928	.957	.927	.959
$4M_{14}$	.946	.95	.944	.948	.932	.939	.94	.946
$4M_{15}$	.949	.954	.937	.953	.899	.949	.903	.952
$4M_{16}$	.956	.95	.947	.949	.909	.94	.913	.936
$4M_{17}$	.944	.959	.933	.957	.921	.957	.929	.957
$4M_{18}$	.948	.955	.933	.953	.918	.955	.911	.948
$4M_{19}$	.943	.959	.943	.958	.931	.956	.933	.956
$4M_{20}$	.948	.951	.943	.949	.928	.944	.938	.948
$4M_{21}$	.948	.946	.943	.95	.924	.937	.925	.94
$4M_{22}$	.941	.952	.935	.948	.932	.946	.936	.945
$4M_{23}$	.961	.959	.957	.966	.929	.954	.938	.957
$4M_{24}$	.968	.956	.968	.958	.933	.948	.947	.958
$4M_{25}$	.948	.955	.949	.953	.914	.943	.892	.945
$4M_{26}$	.945	.955	.949	.961	.931	.949	.926	.954
$4M_{27}$	.958	.952	.948	.962	.94	.949	.932	.955
$4M_{28}$	.948	.947	.951	.961	.924	.944	.923	.956
$4M_{29}$	.947	.946	.96	.954	.925	.947	.929	.952
$4M_{30}$	.952	.942	.959	.944	.923	.946	.888	.931
$4M_{31}$	.947	.949	.946	.943	.911	.95	.895	.94
$4M_{32}$	.944	.959	.94	.958	.927	.951	.936	.953
$4M_{33}$	.944	.955	.94	.958	.93	.952	.941	.952
$4M_{34}$	.942	.952	.949	.953	.932	.947	.936	.953
$4M_{35}$	.935	.951	.951	.956	.925	.945	.94	.95
$4M_{36}$	.935	.945	.952	.948	.918	.932	.893	.936
ALL	.937	.947	.927	.949	.901	.943	.890	.943

Table 10

Type I error rates and Power to detect interaction effect.

	Type I error ($\gamma_3 = 0$)					Power ($\gamma_3 = 0.15$)				
	Non-normal		Normal			Non-normal		Normal		
n	200	926	200	926	Overall	200	926	200	926	Overall
$3M_1$	4.4	3.6	4.4	3.8	4.1	40.2	99.3	48.7	99.8	72.5
$3M_2$	4.3	4.6	5.3	4.8	4.8	48.0	99.6	54.2	99.9	75.5
$3M_3$	3.3	3.8	4.1	4.0	3.8	38.6	98.5	42.9	99.6	70.4
$3M_4$	3.3	3.0	4.8	5.4	4.1	32.3	82.7	33.4	83.9	60.1
$3M_5$	4.2	4.9	6.0	5.1	5.1	50.1	99.7	53.0	99.9	75.7
$3M_6$	3.6	2.4	4.6	5.0	3.8	36.5	89.3	38.4	89.1	65
$3M_7$	4.9	4.4	6.2	4.4	5	54.9	99.5	55.3	99.9	77.9
$3M_8$	5.4	4.2	6.0	4.4	5	55.0	99.5	56.7	99.9	77.9
$3M_9$	5.4	5.4	5.8	4.9	5.4	42.5	98.3	44.3	99.1	71.5
$3M_{10}$	2.9	4.5	3.4	4.9	3.9	23.8	81.0	25.3	81.8	56.2
$3M_{11}$	4.8	5.0	6.1	5.0	5.2	46.8	99.0	49.3	99.6	73.8
$3M_{12}$	2.6	4.6	4.2	4.6	4	29.1	87.7	31.1	88.4	62.3
$3M_{13}$	4.4	4.5	4.5	3.8	4.3	46.6	99.4	45.0	99.7	73.3
$3M_{14}$	2.9	4.0	2.9	4.2	3.5	44.7	98.4	40.0	98.8	72.6
$3M_{15}$	4.2	5.4	3.3	4.6	4.4	37.4	97.9	36.3	98.9	68.1
$3M_{16}$	1.9	1.6	1.9	2.8	2	13.5	75.2	15.6	86.3	51.1
$3M_{17}$	3.3	4.8	2.8	4.6	3.9	28.9	93.9	26.6	93.6	63.3
$3M_{18}$	1.4	2.0	1.0	2.4	1.7	10.9	60.4	11.0	72.9	42.3
$3M_{19}$	5.3	4.4	6.0	4.4	5	56.4	99.4	56.1	99.7	78
$3M_{20}$	4.0	5.2	4.7	4.1	4.5	54.7	99.1	52.5	99.1	77.7
$3M_{21}$	5.5	5.1	4.5	4.6	4.9	47.3	99.0	50.6	99.7	74.2
$3M_{22}$	2.2	3.1	3.1	3.5	3	22.1	80.6	29.1	89.7	58.3
$3M_{23}$	5.8	5.3	3.5	4.9	4.9	37.9	94.1	39.7	93.7	68.3
$3M_{24}$	2.2	3.2	3.1	3.7	3	16.3	67.4	21.8	79.1	49.6
$4M_1$	3.5	2.5	4.7	5.5	4	30.6	82.7	32.4	83.0	59
$4M_2$	3.6	2.8	5.3	4.6	4	37.3	89.2	37.1	88.9	64.6
$4M_3$	4.3	4.7	5.3	5.1	4.9	48.0	99.7	53.1	100.0	75.3
$4M_4$	4.2	5.2	6.5	5.4	5.3	49.3	99.7	52.7	100.0	75.5
$4M_5$	4.3	3.8	4.6	4.0	4.2	40.0	99.3	48.0	99.8	72.2
$4M_6$	3.0	3.8	4.0	4.3	3.8	38.4	98.6	42.8	99.6	70.4
$4M_7$	4.3	4.6	5.2	4.8	4.7	48.8	99.7	55.5	99.9	76.1
$4M_8$	4.1	4.0	5.7	4.0	4.4	47.5	99.5	52.4	99.9	75.2
$4M_9$	4.4	4.6	5.4	4.9	4.8	49.5	99.7	55.3	100.0	76.3
$4M_{10}$	4.3	4.8	5.8	5.2	5	49.7	99.7	54.5	100.0	76.1
$4M_{11}$	4.4	3.6	4.6	4.3	4.2	40.6	98.5	45.8	99.6	71.6
$4M_{12}$	4.2	4.8	5.9	5.2	5	50.2	99.7	54.3	100.0	76.1
$4M_{13}$	5.2	4.9	6.0	4.8	5.2	55.4	99.5	57.2	99.9	78.1
$4M_{14}$	5.1	5.0	5.6	5.2	5.2	47.0	98.9	50.2	99.5	74
$4M_{15}$	5.0	4.6	6.3	4.7	5.1	54.4	99.6	55.2	99.9	77.5
$4M_{16}$	4.2	5.0	5.1	5.1	4.9	42.8	98.6	45.1	99.2	71.7
$4M_{17}$	5.5	4.1	6.7	4.3	5.2	55.0	99.5	55.9	100.0	77.7
$4M_{18}$	5.0	4.5	6.7	4.7	5.2	55.6	99.6	56.9	99.9	78.3
$4M_{19}$	5.6	4.1	5.7	4.2	4.9	55.0	99.5	57.1	99.9	78
$4M_{20}$	4.9	4.9	5.7	5.1	5.2	46.9	98.9	49.2	99.6	73.7
$4M_{21}$	4.9	5.4	5.6	5.0	5.2	44.6	98.3	45.3	99.2	72.2
$4M_{22}$	5.4	4.8	6.5	5.2	5.5	46.1	99.0	49.6	99.6	73.7
$4M_{23}$	3.8	4.1	4.2	3.4	3.9	44.4	99.5	42.6	99.7	71.7
$4M_{24}$	3.1	4.4	3.1	4.2	3.7	36.3	98.6	34.8	99.4	67.4
$4M_{25}$	4.8	4.2	4.1	4.3	4.4	45.2	98.7	42.5	98.8	73.2
$4M_{26}$	4.5	4.5	4.5	3.9	4.3	46.3	99.4	44.9	99.7	73
$4M_{27}$	3.8	4.8	4.6	3.8	4.3	45.9	99.4	44.5	99.8	72.8
$4M_{28}$	4.3	5.3	4.0	3.9	4.4	37.6	98.7	35.8	99.3	68.2
$4M_{29}$	4.8	5.4	3.1	4.6	4.5	38.1	97.9	36.0	98.8	68.1
$4M_{30}$	4.1	5.2	3.0	5.2	4.4	30.7	94.3	29.6	94.7	64.7
$4M_{31}$	4.5	4.8	5.2	4.6	4.8	56.1	99.2	53.7	99.1	78.3
$4M_{32}$	5.2	4.1	6.0	4.2	4.9	56.8	99.4	56.1	99.7	78.1
$4M_{33}$	5.6	4.5	6.0	4.2	5.1	56.2	99.4	56.1	99.7	77.9
$4M_{34}$	5.7	4.8	5.1	4.7	5.1	46.4	98.9	49.6	99.6	73.7
$4M_{35}$	6.4	4.9	4.9	4.4	5.1	47.3	98.9	50.3	99.6	74.1
$4M_{36}$	6.2	5.3	4.5	5.0	5.2	39.3	94.0	40.3	94.3	68.9
ALL	6.1	5.3	7.2	5.1	5.9	55.7	99.7	58.8	100.0	78.6

Table 11

Study 3: Estimating γ_3 with all-pairs and LMS. PARB= parameter bias (absolute for $\gamma_3 = 0$, relative for $\gamma_3 = 0.15$). SEB=Relative standard error bias. COV= coverage rate. RR=rejection rate. D1, D2, D3=distribution conditions.

			ALLPAIRS				LMS			
		n	PARB	SEB	COV	RR	PARB	SEB	COV	RR
$\gamma_3 = 0$	D1	100	0.00	-0.12	92.00	8.00	0.00	-0.01	93.00	7.00
		200	0.01	-0.02	93.00	7.00	0.01	0.05	93.00	7.00
		400	0.00	0.03	94.00	6.00	0.00	0.05	93.00	7.00
		926	0.00	0.05	94.00	6.00	0.00	0.07	94.00	6.00
	D2	100	0.00	-0.14	92.00	8.00	0.00	-0.03	92.00	8.00
		200	0.00	0.02	95.00	5.00	0.00	0.07	95.00	6.00
		400	0.00	0.04	94.00	6.00	0.00	0.04	94.00	6.00
		926	0.00	0.07	94.00	6.00	0.00	0.08	94.00	6.00
	D3	100	-0.01	-0.37	90.00	10.00	0.01	-0.06	90.00	10.00
		200	0.00	-0.15	91.00	9.00	0.01	-0.06	91.00	9.00
		400	0.00	-0.03	95.00	5.00	0.01	-0.03	92.00	7.00
		926	0.00	-0.01	94.00	6.00	0.01	-0.02	91.00	9.00
$\gamma_3 = 0.15$	D1	100	-0.02	-0.14	88.00	35.00	-0.01	-0.01	92.00	37.00
		200	0.01	-0.07	89.00	60.00	0.04	0.03	94.00	64.00
		400	0.00	0.03	93.00	88.00	0.00	0.06	94.00	90.00
		926	-0.01	0.06	94.00	100.00	0.00	0.06	95.00	100.00
	D2	100	-0.03	-0.17	87.00	37.00	-0.02	-0.03	92.00	38.00
		200	0.00	-0.04	91.00	58.00	0.02	0.06	95.00	62.00
		400	0.00	0.02	92.00	89.00	-0.01	0.03	94.00	88.00
		926	-0.01	0.06	95.00	100.00	-0.02	0.08	94.00	100.00
	D3	100	-0.05	-0.32	83.00	33.00	0.00	-0.06	91.00	36.00
		200	0.02	-0.18	85.00	52.00	0.06	-0.09	89.00	61.00
		400	-0.01	-0.10	89.00	74.00	0.05	0.00	92.00	86.00
		926	-0.01	-0.08	91.00	98.00	0.06	-0.05	91.00	100.00

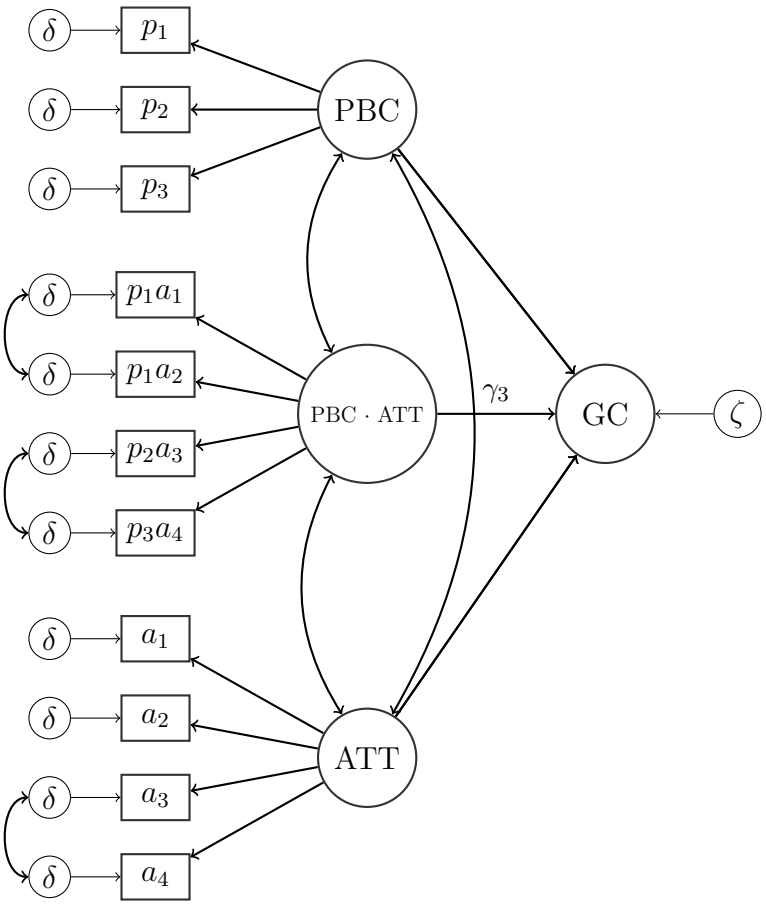

Figure 1. Interaction model for a matched-pair choice of PIs. Measurement model for GC not shown. PBC= Perceived Behavioral Control ; ATT= Attitude; GC= Goal Commitment; Indices for measurement residuals δ not shown; ζ is a regression residual; x_i and x_j are indicators for Perceived Behavioral Control and Attitude, respectively.

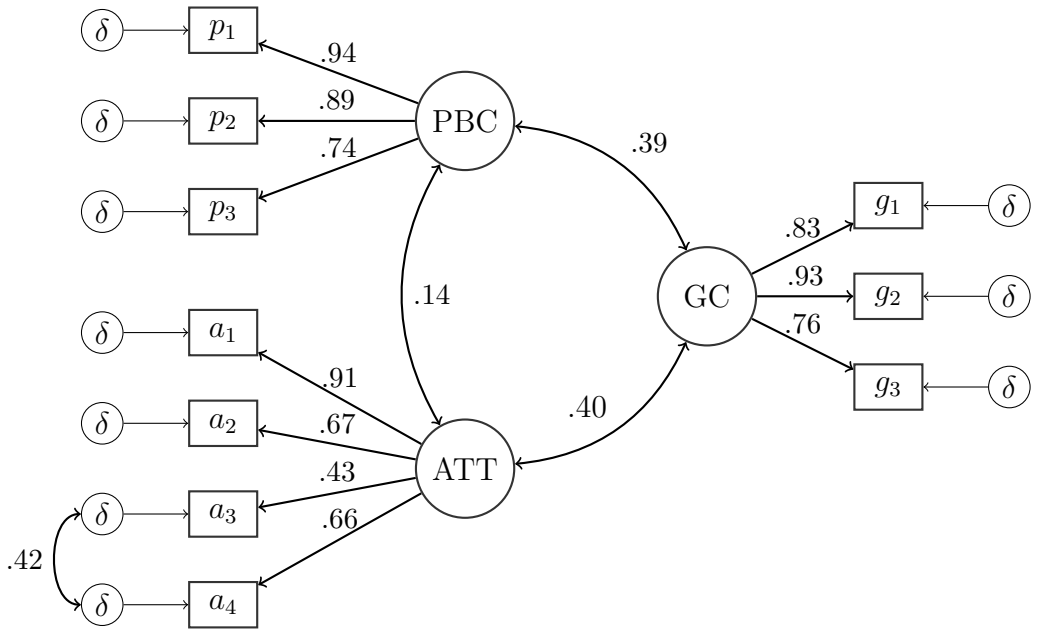

Figure 2. Measurement model with standardized estimates. PBC= Perceived Behavioral Control ; ATT= Attitude; GC= Goal Commitment; Indices for measurement residuals δ are not shown.

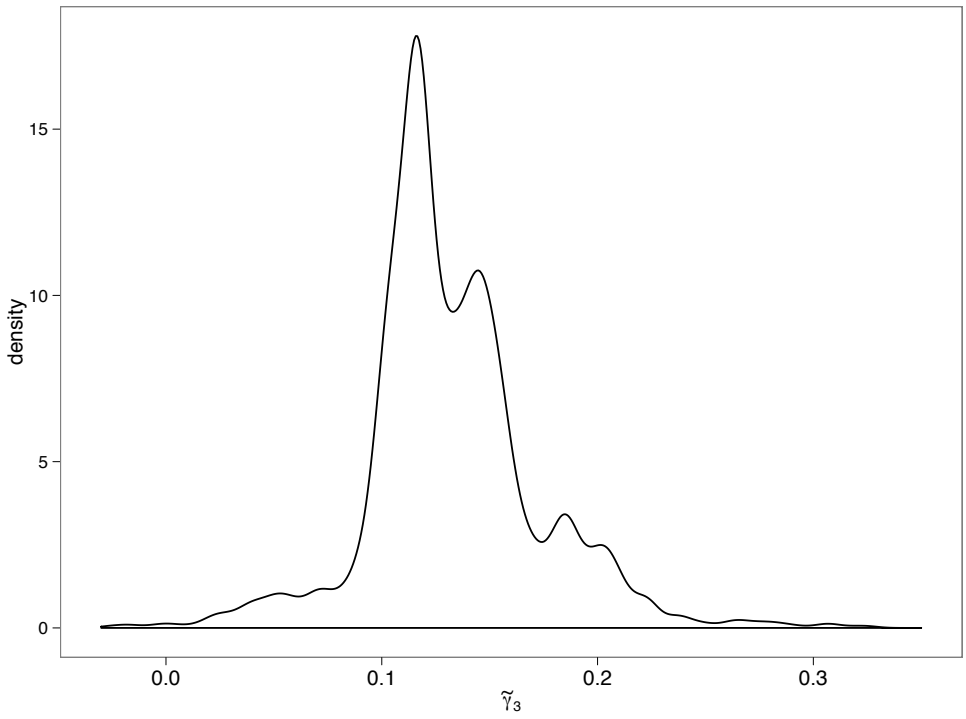

Figure 3. Kernel density plot of the standardized $\tilde{\gamma}_3$ across all converged models.

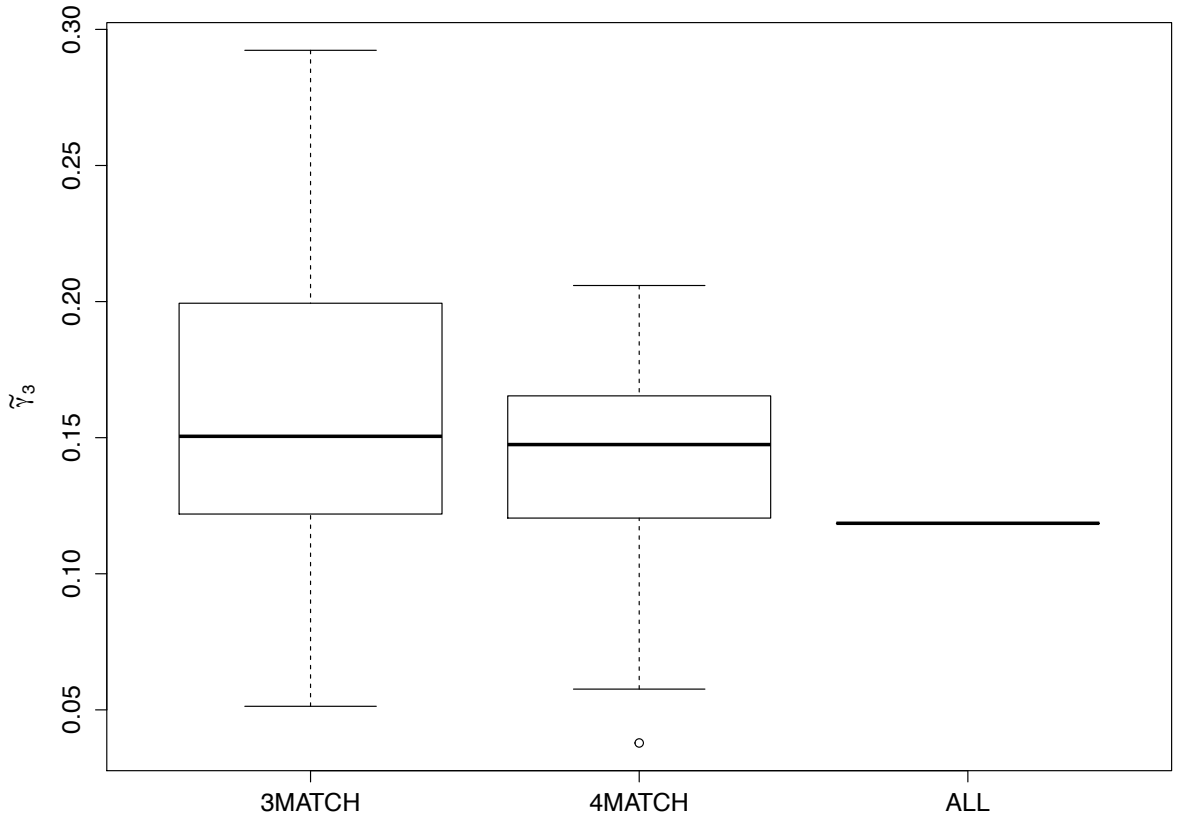

Figure 4. Boxplots of the standardized $\tilde{\gamma}_3$ across 3MATCH, 4MATCH and ALL.

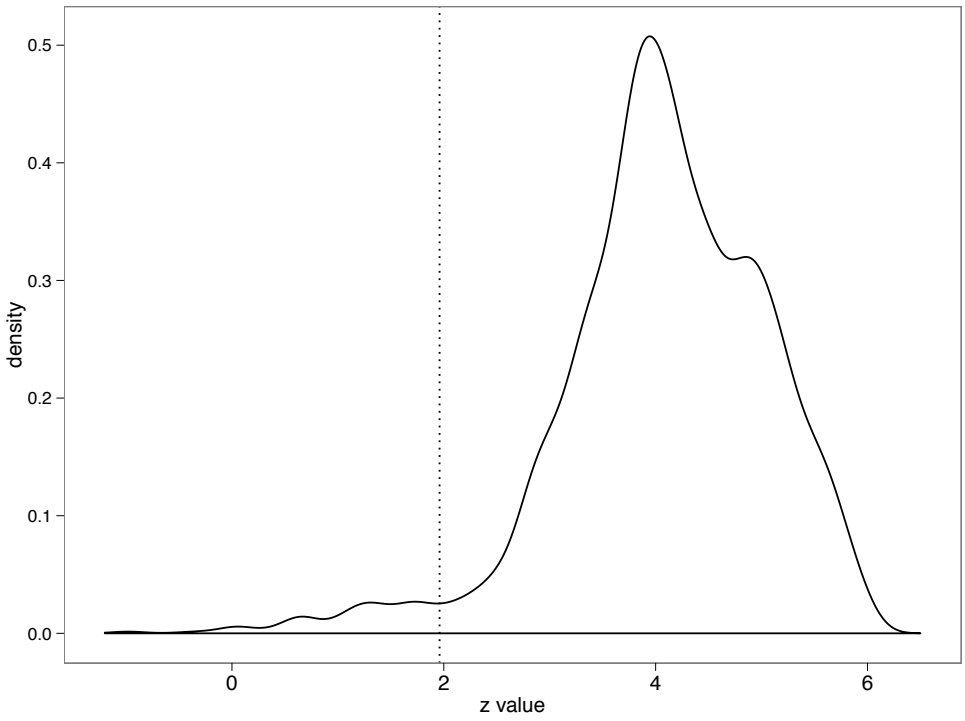

Figure 5. Kernel density plot of the z-value of γ_3 across all models. Dotted vertical line at $z = 1.96$

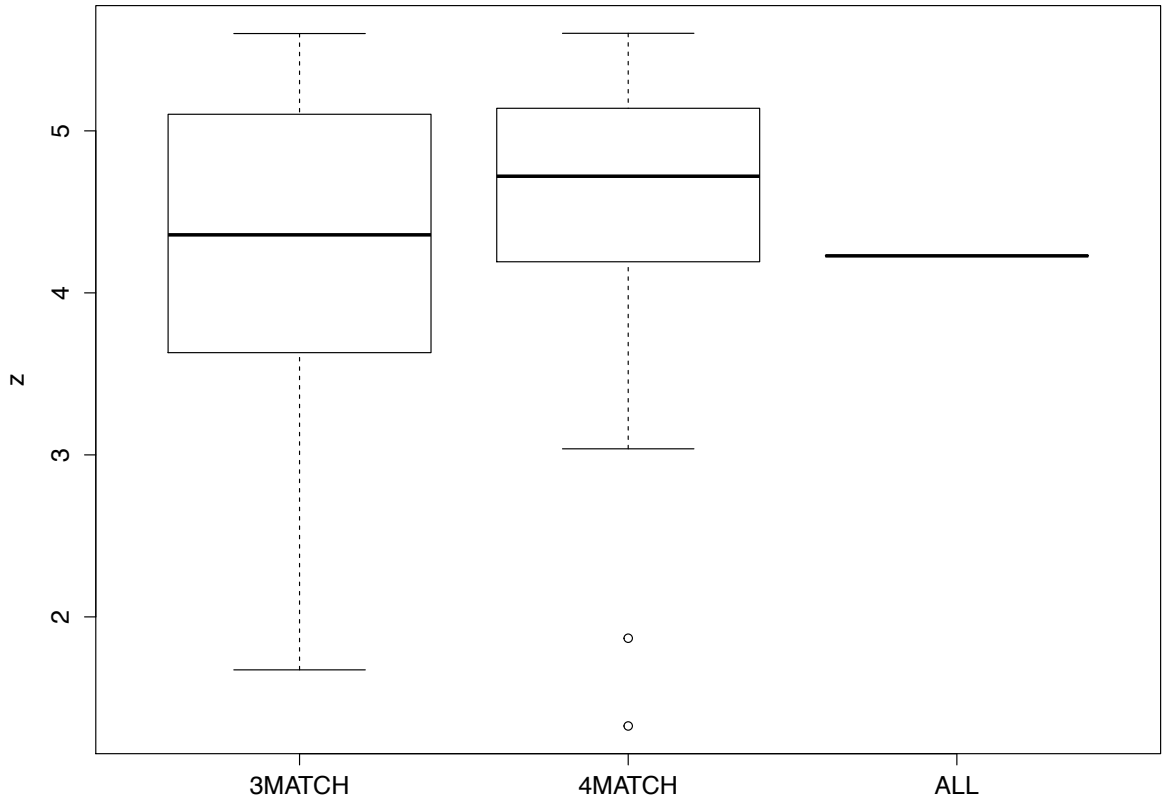

Figure 6. Boxplots of z across 3MATCH, 4MATCH and ALL.

Appendix

Measures

Attitudes (ATT) toward quitting smoking were assessed using four items, denoted by a_1 to a_4. The statement "For me quitting smoking would be..." was completed using the semantic differentials a) wrong–right, b) foolish–wise, c) unpleasant–pleasant, d) unsatisfying–satisfying. The items were scored on a 7-point scale.

PBC was assessed using three items: a) "How much control do you have over quitting smoking?" , b) "How confident are you that you will quit smoking?" , and c) "How certain are you that you are able to quit smoking?" . These items, named p_1, p_2 and p_3 were rated on a 7-point scale ranging from 1 (no control) to 7 (much control).

The measure of goal commitment (GC) was used as a proxy for intention and measured by three items, denoted by g_1, g_2 and g_3. These are: During the next four months: "I have made plans when to quit smoking" , "I have made plans how I am going to quit smoking" , and "I have made plans regarding what I am going to do when temptation situations arise" . The items were rated from 1 (completely wrong) to 7 (completely correct). For further information the reader is referred to Luszczynska and Schwarzer (2003); Hukkelberg et al. (2013).