Social Indicators Research
https://doi.org/10.1007/s11205-025-03584-z

ORIGINAL RESEARCH

Detecting Causal Relations Among Indicators with the CTA Test: Simulations and Applications

Mattia Cefis[1] · **Mario Angelelli**[2] · **Maurizio Carpita**[1] · **Enrico Ciavolino**[2]

Accepted: 13 March 2025

Abstract
In the context of using structural equation modelling to develop economic and social indicators, a debate regarding the choice of measurement modes for theoretical constructs is becoming a very important issue, with conceptual and practical implications. The nature of each construct, which can be defined as reflective or formative, is mainly based on theoretical considerations, but confirmatory tetrad analysis (CTA) can support decisions about the model specification. One flexible approach to carrying out CTA involves multiple hypothesis testing, which also provides relevant information on empirical data to guide the construction of composite indicators. This prompts a deeper investigation of the effects of correction methods on decisions derived from tests, with special attention to error control and statistical power. In this study, we explore the properties of six procedures, in particular the well-known Bonferroni and Benjamini–Hochberg corrections, using various simulation scenarios and real applications. We find that, with respect to the Benjamini–Hochberg, the Bonferroni correction is too conservative and has lower power, especially with small sample sizes and many manifest variables.

Keywords CTA · Multiple hypothesis testing · Measurement mode · PLS-SEM · Bonferroni correction · Benjamini–Hochberg correction

Mario Angelelli, Maurizio Carpita, and Enrico Ciavolino contributed equally to this work.

✉ Mattia Cefis
mattia.cefis@unibs.it

Mario Angelelli
mario.angelelli@unisalento.it

Maurizio Carpita
maurizio.carpita@unibs.it

Enrico Ciavolino
enrico.ciavolino@unisalento.it

[1] Department of Economics and Management, University of Brescia, Contrada Santa Chiara, 50, 25121 Brescia, Italy

[2] Department of Human and Social Sciences, University of Salento and CAMPI, Via di Valesio, 24, 73100 Lecce, Italy

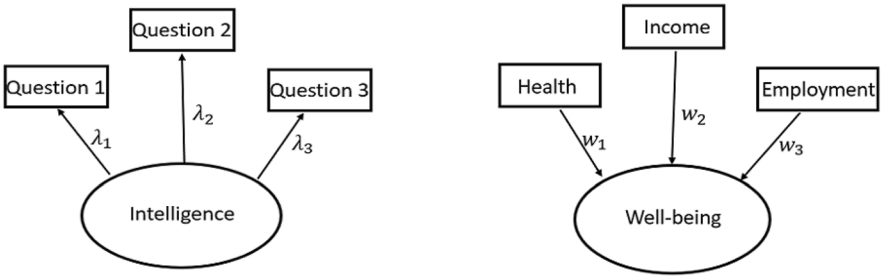
Published online: 10 April 2025

1 Introduction

In the domain of structural equation modelling (SEM) for developing economic and social indicators, the debate over selecting appropriate measurement modes for theoretical constructs has become a significant issue. This discussion holds both conceptual and practical significance. In particular, there are different approaches, for example partial least squares structural equation model (PLS-SEM) (Wold, 1985), an increasingly used tool to define and measure theoretical multidimensional constructs with latent variables (LVs) starting from some manifest variables (MVs); then, the classical covariance-based SEM (CB-SEM) (Joreskog, 1973) and another statistical model-based approach (Cavicchia & Vichi, 2021) for the construction of composite indicators with hierarchical structures. Many recent papers have pointed their attention to this issue, from its practical approach to the methodological one (Cheah et al., 2019, 2021; Ciavolino et al., 2022; Sarstedt et al., 2016).

The nature of each construct, which can be defined as reflective or formative, is mainly based on theoretical considerations, but confirmatory tetrad analysis (CTA) can support decisions about the model specification. One flexible approach to carrying out CTA involves multiple hypothesis testing, which also provides relevant information on empirical data to guide the construction of composite indicators. In particular, in a reflective construct causality is directed from the latent variable (LV) to items (Manifest Variables, MVs), while the formative construct implies that the LV is determined as a combination of its own MVs (i.e. causality from items to construct) (Bollen & Ting, 2000).

A typical example of a reflective model is the case of an intelligence test (Fig. 1a), where each block of MVs (in this case, each question of the test) reflects its LV. Note that reflective indicators are interchangeable; in fact, if we remove one item, we do not alter the underlying concept. Reflective models also assume uni-dimensionality for each block of MVs (just one latent concept is reflected on different indicators), and the loadings λ_i can be estimated through ordinary least squares (OLS). On the other hand, there may be theoretical or conceptual reasons to consider a block as formative, starting with a strong consensus among experts about how the latent variable is formed. For example, in Fig. 1b we can see well-being as LV caused by its own MVs (i.e., health, income, and employment). Compared with the reflective model, in this case, if we omit one MV, we lose a part of the concept (Coltman et al., 2008); in fact, formative indicators are not supposed to be correlated, and for this reason they cannot be evaluated in the same way



(a) Reflective: the Intelligence Index. (b) Formative: the Well-Being index.

Fig. 1 Examples of reflective and formative models

as reflective measures. Each LV is considered to be formed by its MVs following a multiple regression, and weights w_i are estimated by least squares methods.

So, the debate between reflective and formative constructs is open, and it is treated in different papers: Simonetto (2012) gave a panoramic view by a bibliometric analysis about this topic; in addition, several tests exist to assess reflective models, even in the presence of higher-order constructs (Rajala & Westerlund, 2010; Ingusci et al., 2024; Ciavolino et al., 2024, 2022) or additional structure, such as external information (Ciavolino et al., 2015) or longitudinal effects (Ingusci et al., 2023). On the other hand, researchers mainly rely on theory and experts' opinions for formative ones, as the classical procedures to assess reflective constructs are not suitable for the formative ones (Diamantopoulos & Winklhofer, 2001; Chang et al., 2016). The distinction between the two measurement modes has both methodological and practical implications. First, the two models rely on different premises, including the aforementioned unidimensionality and correlation assumptions for reflective models, and they require different evaluation approaches. In particular, validity measures for reflective models, such as average variance extracted (AVE) and reliability indices, are not well-defined in formative models (Hair et al., 2020, Sec. 3.2), where no measurement error is associated with formative indicators (Diamantopoulos et al., 2008). As a consequence, such indices targeted on reflective models may improperly suggest low performance for valid but formative indicators (Bollen & Lennox, 1991). More importantly, these problems can cause measurement model misspecification; as a consequence, in PLS-SEM, they can lead to biases in the inner model estimation and an incorrect assessment of relationships (Diamantopoulos et al., 2008, Sec. 4) [also see Bollen (2002, p. 616)]. These measurement approaches reflect different item independence structures conditioned on the latent variable, as occurs in a broader class of causal modelling and latent trait frameworks such as Item Response Theory (Bollen, 2002). In psychological and social research, distinctions between such models are often unsharp, making the choice dependent on the study's conceptualisation and objectives [see, e.g., Sarstedt et al. (2016, Sec. 2.2) and references therein]. Our contribution aims to support researchers in this selection process, particularly when the lack of strong theoretical motivations places greater emphasis on explanatory and predictive objectives; notably, this focus is also a primary rationale for adopting PLS-SEM (Hair et al., 2019).

In order to overcome those limitations, some researchers have applied confirmatory tetrad analysis (CTA) (Bollen & Ting, 1993), a statistical support to draw conclusions about the appropriateness of using formative measurement models as compared to reflective ones (Gudergan et al., 2008). For a case study in line with the example provided in Fig. 1, see, e.g., Bollen et al. (2009).

The CTA was initially proposed for the CB-SEM models: the original CTA procedure uses a multiple hypothesis testing approach, and Bollen and Ting (1998) developed a bootstrapping procedure for computing the p value of the CTA test statistic in order to have greater accuracy than using the original chi-square distribution, examining the test by means of simulations and empirical examples. The final version of the CTA applied to the covariance-based models (CTA-SEM) was refined by those authors in a test for causal indicators (Bollen & Ting, 2000). The equivalent procedure in the context of PLS-SEM was developed by Gudergan et al. (2008), starting from the classical CTA; the authors proposed a CTA-PLS assessment routine for measurement models in order to be consistent with PLS assumptions, with a validation performed by a bootstrapping procedure. CTA-PLS aims to distinguish formative indicators from reflective ones and is well-implemented into the SmartPLS software (Ringle et al., 2015).

The information provided by CTA is independent of the chosen estimation method (CB-SEM or PLS-SEM) and the analysis objective (confirmatory methods based on fitting or variance-based methods focusing on predictive performance). In fact, CTA extracts information based on the empirical covariance matrix, which can guide the measurement model specification regardless of the estimation methods for loadings, scores, and path coefficients in the structural model. The difference between an omnibus test (Bollen & Ting, 2000) and multiple testing (Gudergan et al., 2008) regards the information extracted to support a measurement mode compared to the alternative; finer information emerges from multiple hypothesis testing, but this advantage should be balanced by a careful analysis of Type I errors and statistical power. Furthermore, the CTA-PLS approach does not rely on distributional assumptions, which makes it suitable for different analyses with specific estimation procedures and objectives (confirmation, explanation, or prediction). This work addresses this issue and investigates the effect that different correction methods can have on measurement mode selection based on statistical tests.

In particular, CTA-PLS was applied in a few empirical examples in order to test the nature of some LVs: Tabet et al. (2020a) used it firstly for an analysis of the World Health Organisation disability assessment schedule, then for analysing the factor structure of the outcome questionnaire (Tabet et al., 2020b), while Sarstedt et al. (2021) provided some empirical examples in the well-known book of PLS-SEM. In addition, Cefis and Carpita (2022) applied the CTA-PLS to test the nature of some football performance indicators based on some experts' opinions. More recent papers adopted CTA-PLS for supporting experts in some policy government decisions: Mendes et al. (2023) explored how innovation activities and cluster affiliation moderate the relationship between family involvement and post-internationalisation speed in family firms, while Ongena (2023) proposed data literacy for improving governmental performance. This work expands the current literature by analysing both simulated and empirical data to compare different correction methods, discussing in detail the factors (sample size, number of manifest variables) that may affect the performance of multiple testing in CTA and the subsequent decisions regarding the nature of latent variables.

The paper is organised as follows: in Sect. 2 the CTA-PLS multiple hypothesis test is described, Sect. 3 shows the simulation framework and discusses the obtained results. In Sect. 4, some empirical applications are presented, and final conclusions are provided in Sect. 5.

2 Multiple Testing Corrections in CTA

2.1 The CTA-PLS Test

The Confirmatory Tetrad Analysis proposed in the context of the PLS-SEM (CTA-PLS) is a statistical tool based on a confirmatory approach for the evaluation of cause-effect relationships (i.e. reflective or formative) in the measurement models (Gudergan et al., 2008). This technique is relevant for a posterior re-examination of the constructs to assess possible misspecifications of the measurement models. In fact, researchers recommend a priori theoretical specification and posterior re-examination along with empirical data, which are essential to better understand the structure of the outer models (Gudergan et al., 2008). From a practical point of view, after setting all the measurement models in a reflective way, a CTA-PLS can be applied in order to understand which LVs are confirmed as reflective

constructs and which ones are not (Tabet et al., 2020a); then, PLS-SEM can be applied using the type of constructs provided by the CTA-PLS. We stress that, at this stage, the procedure places CTA before model estimation and, hence, does not rely on the estimation method, as its inputs are the measurement models and the empirical covariance matrix.

In the CTA-PLS context, we must define the concept of a generic tetrad τ: it is the difference between the product of two pairs of covariances. For instance, the six covariances of a block of 4 MVs involve two non-redundant vanishing (i.e. equal to zero) tetrads:

$$\begin{aligned} \tau_{1234} &= \sigma_{12}\sigma_{34} - \sigma_{13}\sigma_{24} = 0 \\ \tau_{1243} &= \sigma_{12}\sigma_{43} - \sigma_{14}\sigma_{23} = 0 \end{aligned} \tag{1}$$

It should be taken into consideration that the construction of tetrads in (1) requires 4 MVs per time, while the original CTA is also applicable to measurement models with a different number of indicators (Bollen & Ting, 1993). In fact, the CTA-SEM originally introduced in the CB-SEM applications (Bollen & Ting, 2000) presented the concept of vanishing tetrads using a covariance data matrix to complement standard procedures of model evaluation and provided methods for selecting model-implied non-redundant vanishing tetrads and significance testing. In the general case of J non-redundant tetrads, the following multiple hypothesis testing is considered:

$$\begin{aligned} H_0 &: \quad \text{all } H_0^{(j)} : \tau_j = 0 \text{ for } j = 1, 2, ..., J \\ H_1 &: \text{ at least } H_1^{(j)} : \tau_j \neq 0 \text{ for } j = 1, 2, ..., J \end{aligned} \tag{2}$$

By (2), if the test does not reject H_0, the construct can be confirmed as reflective from the statistical point of view; otherwise, if at least one tetrad j of the block does not vanish, the test suggests a formative measurement model.

Although CTA-PLS uses a similar evaluation process, the approach differs from CTA-SEM for PLS methodological assumptions in both the single tetrad testing approach and the simultaneous tetrad testing procedure. In fact, CTA-PLS builds on the statistical test for every single measurement model-implied vanishing tetrad. To overcome the limitations regarding distributional assumptions, it includes a bootstrapping routine (Gudergan et al., 2008). It should be taken into account that neither CTA-SEM nor CTA-PLS are applicable for covariances of MVs close to zero in the measurement model (Bollen & Ting, 2000) if some key MVs of a block are uncorrelated; for example, if $\sigma_{12} = \sigma_{13} = \sigma_{14} = 0$ in (1), all tetrads would by definition equal zero, which makes the CTA-PLS meaningless. Therefore, it first requires testing whether at least some of the measurement model's indicators are significantly correlated (Hair et al., 2017). A potential effect of the estimation method may emerge for an extended set of model-implied tetrads derived from LVs with less than 4 indicators, e.g. based on Gudergan et al. (2008, Table 1) and specified criteria (Cheah et al., 2019; Puche-Regaliza et al., 2021).

Since in CTA-PLS all tests are made for all the non-redundant tetrads in each block of MVs, a multiple testing problem is involved. In order to deal with this issue, a Bonferroni adjustment of the significance levels is used. It assures that the error rate does not exceed the level α for all the J desired tests. The Bonferroni approach lets us compute simultaneous statistical test values for multiple tetrad tests; in contrast to the CTA-SEM (Bollen & Ting, 1998), in which an asymptotic chi-squared test is performed, a t test is adopted by the CTA-PLS (Gudergan et al., 2008).

The practical application of CTA-PLS is similar to that of CTA-SEM (Bollen & Ting, 2000):

1. All tetrads for the measurement model of a given LV are computed. In general, given m MVs for one LV, the binomial coefficient

$$C_{m,4} = \frac{m!}{(m-4)! \cdot 4!} \tag{3}$$

computes the number of sets of 4 variables, each resulting in 3 vanishing tetrads for measurement models with m MVs for each block. So, the total number of tetrads for each one is:

$$\#\tau = 3 \cdot C_{m,4} \tag{4}$$

Despite CTA-SEM, it is interesting to highlight that a LV with $m < 4$ requires the inclusion of indicators from another one to form a set of four MVs to perform CTA-PLS.

2. In this step, the tetrads are identified. For optimising the definition of H_0 and computational time, a minimal set of tetrads is selected, whose vanishing entails that the redundant tetrads equal 0 too. This kind of redundancy arises from algebraic relations or when the same pair of covariances appears in two tetrads.
3. A statistical significance test for each tetrad is performed, checking whether the value is significantly different from zero. CTA-PLS follows, as introduced before, some of Bollen's suggestions, like using a bootstrap routine (Bollen & Ting, 1998). The generation of a high number of bootstrap subsamples and computing their relevant tetrads allows for obtaining the bootstrap estimated standard error (SE_b) for each one and, then, the observed t statistics:

$$t_j = t_j^* / SE_b(t_j^*) \tag{5}$$

with t_j^* the tetrad estimate of the corresponding τ_j in (2).

4. In the last step, CTA-PLS evaluates the results for all model-implied non-redundant vanishing tetrads by accounting for multiple testing issues. A reflective measurement model does not meet the empirical data if at least one of the model-implied vanishing tetrads is significantly different from zero. CTA-PLS employs a procedure for testing J single null hypotheses (non-redundant vanishing tetrads) $H_0^{(1)}, H_0^{(2)}, \ldots, H_0^{(J)}$ with test statistics $t_1, t_2, \ldots, t_J$ of type (5) for each LV. Thus, in order to take into account multiple testing issues, the rejection probability of the full null hypothesis H_0 in (2) requires adjustment (Gudergan et al., 2008). The default correction used by CTA-PLS is the Bonferroni adjustment, which consists of rejecting $H_0^{(j)}$ if the associated statistic test t_j is significant at the $\alpha' = \alpha/J$ adjusted level of the test, where J is the number of hypotheses to be tested. For instance, if we have 5 non-redundant vanishing tetrads to test and $\alpha = 0.05$, then for each individual test we should use a critical value α' equal to $0.05/5 = 0.01$.

2.2 Multiple Testing Corrections

As explained in the previous section, the CTA-PLS test provided by Gudergan et al. (2008) and available in the SmartPLS software involves a multiple-test issue: it adopts the classical Bonferroni correction (that we consider a benchmark) for adjusting p values but also suggests the possibility to deal with other ones, such as Benjamini and Hochberg (1995). We focused attention on the families of these two approaches:

- Familywise Error Rate (FWER): this family was defined as the probability of incorrectly rejecting at least one $H_0^{(j)}$. FWER is believed to be too conservative in cases where the number of simultaneously tested hypotheses reaches several hundreds or thousands, or when hypotheses are highly correlated (as in the case of the Bonferroni correction used in the CTA-PLS). Given J hypotheses to be tested and a fixed α, the Bonferroni correction (the same approach described at step 4 of the previous page) computes the adjusted p value for each hypothesis $p\text{-}value_j^{Bo} = p\text{-}value_j \cdot J$. The other FWER corrections considered in this study were proposed by Holm, Hochberg, and Hommel. On the basis of Bonferroni correction, Holm (1979) computes the significance levels depending on the p values based-rank in ascending order of hypotheses. Similar to the Holm correction, Hochberg (1988) employs the same formula to compute the associated significance levels but uses a descending order for the p values; in addition, the Hochberg adjustment is more powerful than the Holm one. Another FWER correction is provided by Hommel (1988).
- False Discovery Rate (FDR): in this family, we find the Benjamini–Hochberg correction; FDR is defined as the expected proportion of incorrectly rejected $H_0^{(j)}$ among all rejections. Therefore, FDR allows the occurrence of Type I errors below a reasonable proportion by taking the total number of rejections into consideration. An interesting advantage of FDR is the higher power of statistical inference, which would be useful when a large number of hypotheses are simultaneously tested. Benjamini and Hochberg (1995) introduces a procedure for controlling FDR that is less stringent with the increased gain in power and has been widely used in cases where a large number of hypotheses are simultaneously tested. A sequential approach to controlling the FDR in multiple comparisons, due to Benjamini–Hochberg, yields much greater power than the widely used Bonferroni procedure that limits the familywise Type I error rate (Thissen et al., 2002). In particular, given a multiple test with J hypotheses to be tested and a fixed α, the Benjamini–Hochberg procedure computes the adjusted p value for each hypothesis (Benjamini et al., 2009) as follows:

1. Order all p values from the smallest to the largest, multiply each one by the total number of tests J, and divide it by its rank r.
2. Check if the resulting sequence is non-decreasing; if it does not hold, it makes the preceding p value equal to the subsequent (repeatedly, until the whole sequence becomes non-decreasing).
3. If any p value ends up larger than 1, set it equal to 1:

$$p\text{-}value_j^{BH} = min\left(\min_{r \geq j}\left(\frac{J \cdot p\text{-}value_r}{r}\right), 1\right) \tag{6}$$

Another FDR approach is proposed by Benjamini and Yekutieli (2001) (BY), which is similar to Benjamini–Hochberg, but more conservative.

3 Simulation Plan and Results

In this section, we compare the different corrections based on data simulated from a composite-based structural equation model. As remarked in the Introduction, the multiple hypothesis testing and the correction's results do not depend on the model's estimation method; while the following simulation study can be used in combination with other

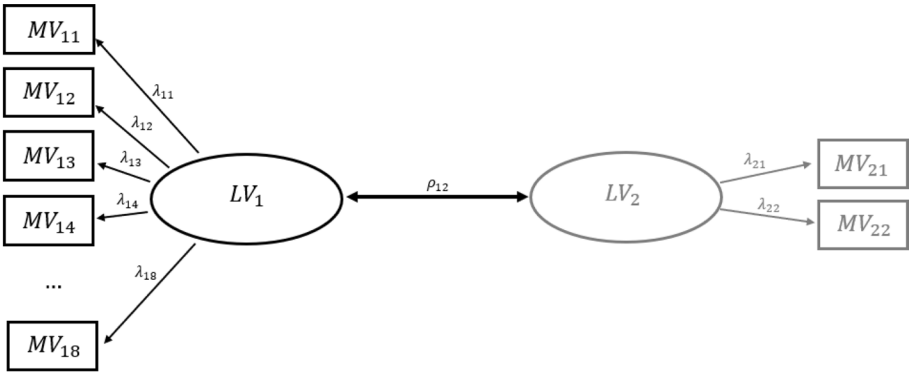
testing approaches for CTA, we discuss it in relation to CTA-PLS as multiple testing naturally emerges in this technique.

In order to assess the performances of the CTA-PLS multiple test corrections described in the previous section, we developed a simulation plan using the R package csem.DGP (version 0.1.0.9000, already used and illustrated by Schamberger (2023)) and the *lavaan* syntax (Schlittgen et al., 2020), already used in other PLS-SEM simulation studies (Danks et al., 2020; Dolce et al., 2022). Following Bollen and Ting (1998), in our simulation, a different number of MVs and sample sizes have been used. Since the CTA-PLS multiple testing is used for each single LV of the SEM, in the case of a single reflective (for the actual significance level) and formative (for the actual power) measurement model of LV_1, we increased the number of MVs from 4 to 8 (i.e. from 2 to 20 tetrads to test); the corrected p values of the multiple tests were computed using the approaches described in the previous section, in particular the Bonferroni and the Benjamini–Hochberg corrections. As the csem.DGP procedure requires at least two LVs in the SEM but the focus was exclusively on LV_1, we used a fictitious LV_2 with correlation $\rho_{12} = 0.2$ (Fig. 2; changing the correlation value, test results do not change). Finally, after some preliminary check on the results' stability, we have chosen to run 40,000 replications for each combination of the number of MVs and the sample sizes, each with 5000 bootstrap resamples, to compute the Standard Error (SE_b) of the statistic t_j in (5).

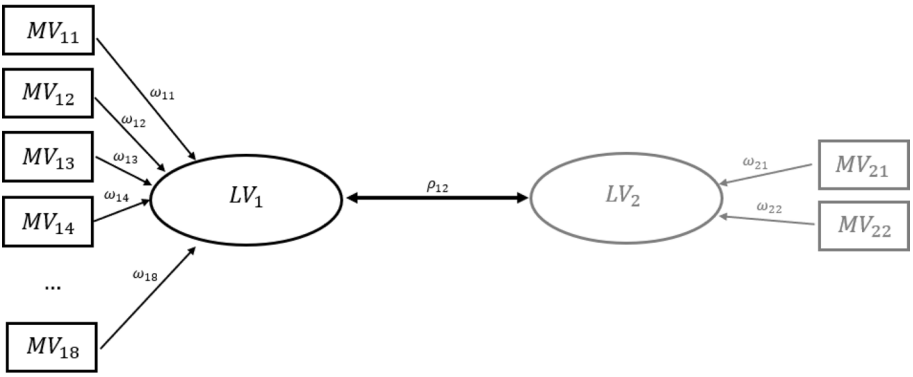
3.1 The Actual Significance Level of the CTA-PLS Multiple Test

We generated data from a reflective structure (Fig. 2a) in order to assess the actual type I error probability of the CTA-PLS multiple tests, i.e. the probability to reject H_0 in (2) when it is true, increasing the number of MVs (from 4 to 8) and the sample sizes ($n = 125, 250, 500, 750, 1000$). The CTA-PLS test Actual Significance Level (ASL) and its 90% confidence interval (CI) for each multiple test correction in Sect. 2.2 were computed, setting the Nominal Significance Level (NSL) at $\alpha = 0.1$. For the reflective case, we chose to adopt equal loadings as for the classical parallel-test model (Nunnally, 1994) and the single-item reliability minimum value of 0.7 (Gudergan et al., 2008). Figure 3 shows the ASLs and their 90% CIs for the multiple test corrections (Bonferroni and Benjamini–Hochberg, see Sect. 2.2), sample sizes, and different numbers of MVs' in the measurement models. The ASLs are lower than the NSL by construction, so it is interesting to underline for all the cases a significant difference between Bonferroni and Benjamini–Hochberg performance (the CIs do not overlap): the Bonferroni correction has a lower Type I error probability than the Benjamini–Hochberg one.

Another interesting result of these simulations is the following: increasing the number of MVs in the measurement model, the Type I error probability of the CTA-PLS is much lower than the NSL = 0.1 for small sample sizes, in particular for $n = 125$ and $n = 250$; with 4 MVs and $n = 125$, respect to the fixed NSL, ASL for the Bonferroni correction is around 7.5% and it reaches the 9% only when $n \geq 750$; ASL with the Bonferroni correction reduces a lot when the number of MVs increases, for small sample sizes: when MVs = 8 and $n = 125$, ASL does not exceed the 3.5% (it reaches the 8% only when $n = 1000$). In this simulation framework, a consistent difference is observed (CIs do not overlap in all the cases) between the Bonferroni and Benjamini–Hochberg corrections. From this perspective, practitioners might favor the more conservative Bonferroni method, as it results in a lower type I error rate. We also adopted the other corrections mentioned in Sect. 2.2,



(a) The Reflective SEM-PLS, with all $\lambda_{ij} = 0.7$ and $\rho_{12} = 0.2$.



(b) The Formative SEM-PLS, with all $\omega_i \overset{iid}{\sim} Unif(0.3, 0.6)$ and $\rho_{12} = 0.2$.

Fig. 2 The path diagrams used in the simulation

but their simulated ASLs were equal to the Bonferroni ASL, so we did not include them in Fig. 3.

3.2 The Actual Power Level of the CTA-PLS Multiple Test

We generated data from a formative structure (Fig. 2b) in order to assess the power of the CTA-PLS multiple tests, i.e. its capability to reject H_0 in (2) when it is false. The Actual Power Level (APL) and its 90% CI for each multiple test correction in Sect. 2.2 were computed. For the formative case, there are not as many reference models as in the reflective one; for this reason, we considered random weights ω_{1j} in a reasonable interval (a uniform distribution between 0.3 and 0.6). However, preliminary simulations showed that modifying weights did not change the simulation results. In addition, the formative model structure needs the initial correlation matrix among MVs. To obtain more general results, we did not use a specific correlation matrix but a random correlation matrix with fixed composite reliability for each replication; in particular, taking the cue from an example available in Gudergan et al. (2008), we used a general framework by simulating random

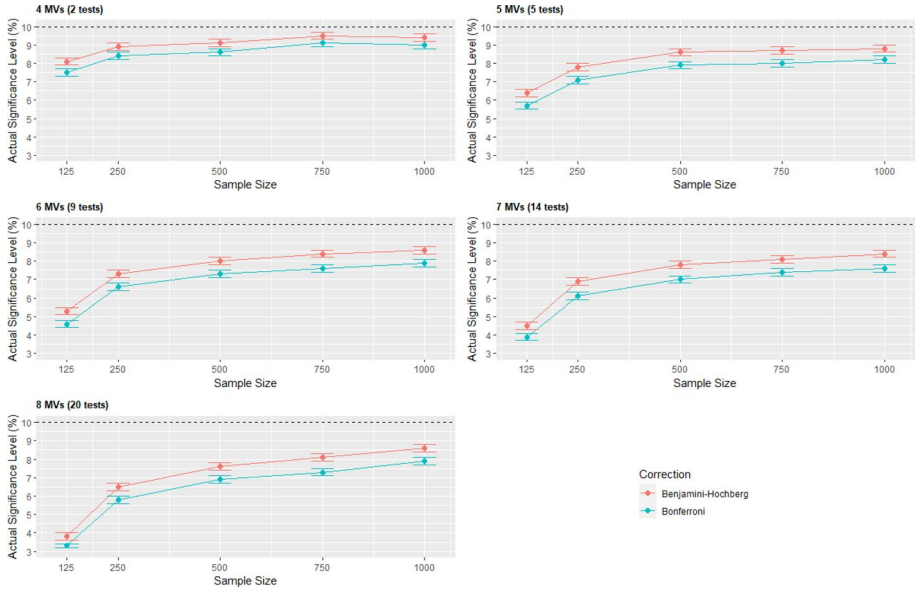


Fig. 3 The CTA-PLS test actual significance level (ASL % with 90% CIs) with 40000 replications and 5000 bootstrap resamples for increasing number of MVs and sample sizes for the reflective model in Fig. 2a

correlation matrices with correlations between 0.1 and 0.5 and a composite reliability index $\rho_c = 0.7$ (see the "Appendix"). For each MV measurement model, we generated data for eight different sample sizes ($n = 125, 200, 250, 300, 400, 500, 750, 1000$) in order to explore the power growth.

In Fig. 4, we show the simulation results for the formative model. In Fig. 2b, we represent the CTA-PLS test actual power level (APL) of the Bonferroni and the Benjamini–Hochberg corrections with their 90% CIs, increasing the sample size for a different number of MVs. Increasing the sample size, for all the measurement models, the APL of the two corrections increases, and a significant difference between them exists (the CIs do not overlap) for sample sizes lower than 500; in particular, the Benjamini–Hochberg correction appears to be more powerful than the Bonferroni one. In addition, by increasing the number of MVs in the measurement model, for small and medium sample sizes, the gap between the two APLs increases too.

For small sample sizes, the two APLs decrease when the number of MVs increases. Considering, for example, the Benjamini–Hochberg correction and $n = 125$, with 4 MVs the APL is 36%, and with 8 MVs, the APL is 22%. APL does not reach the maximum power with the highest sample size ($n = 1000$) for the configurations with 4 and 5 MVs for both approaches (it is lower than 90% with 4 MVs and around 95% with 5 MVs). Based on these results, if the primary focus is on test power, we recommend that practitioners adopt the Benjamini–Hochberg correction, particularly when working with small sample sizes ($n \leq 400$) or complex structures (e.g., measurement models with more than four manifest variables). This approach ensures significantly higher power for multiple testing compared to the Bonferroni correction.

Finally, we have considered the other corrections in Sect. 2.2. We find that the Holm and the Hochberg corrections give exactly the same results as the Bonferroni correction. For the sake of brevity, Fig. 5 shows the APLs of the other four corrections for the case of

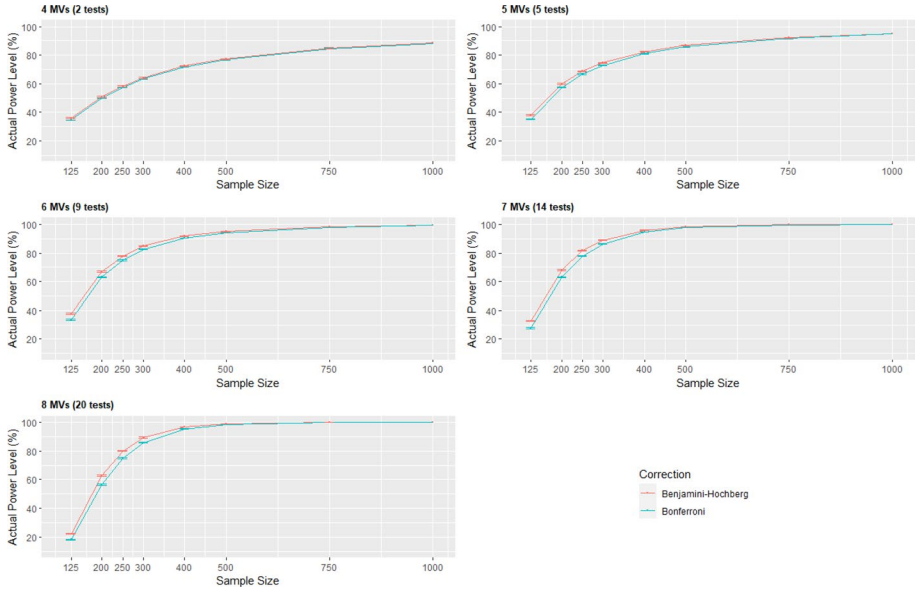


Fig. 4 The CTA-PLS test actual power level (APL %, with 90% CIs) with 40,000 replications and 5000 bootstrap resamples for increasing number of MVs and sample sizes for the formative model in Fig. 2a

8 MVs (20 multiple tests) and a sample size up to $n = 750$. Benjamini–Hochberg is still the more powerful correction, whereas Bonferroni and Hommel corrections perform in a similar way (their CIs do not overlap for small and medium sample sizes). The worst performance (lowest power) is provided by the Benjamini–Yekutieli correction, especially for small and medium sample sizes.

4 Empirical Applications of Multiple Testing in CTA

In this section, we consider real data associated with measurement and structural models to assess the role of the multiple test correction methods in Sect. 2.2 in the study of theoretical constructs through the corresponding LVs. The following case studies are chosen based on three main factors: the accessibility of the data, the association with a measurement model with a suitable number of indicators to conduct CTA-PLS, and the dimensionality of the dataset.

We started by reviewing papers sharing open data and using them with measurement and structural models. Due to our focus on multiple hypothesis testing for measurement models and differences in adjustment methods, we concentrate on models with at least one LV with at least four MVs. Once again, we stress that the subsequent analysis does not rely on the adoption of PLS-SEM as the estimation method, which may only affect the choice of model-implied tetrads for LVs with less than 4 indicators. For this reason, we will distinguish tetrads obtained from LVs with at least 4 indicators from other model-implied tetrads in reporting the CTA-PLS results.

Regarding the size of the dataset, the following analysis involves samples having a number of observations comparable to the simulations conducted in the previous sections.

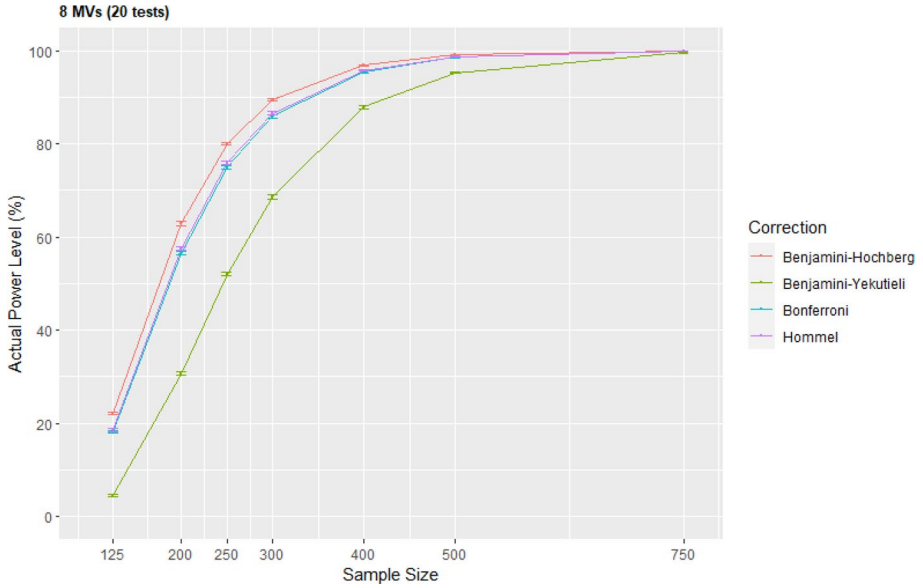


Fig. 5 The CTA-PLS test actual power level (APL %, with 90% CIs) with 40,000 simulations and 5000 bootstrap resamples with 4 multiple test corrections for the formative model in Fig. 3b with 8 MVs

4.1 The ECSI Dataset and Model

The first dataset is the European Customers Satisfaction Index (ECSI) for mobile phone users (2005), with a size of $n = 250$, whose analysis is in continuity with Gudergan et al. (2008). The PLS-SEM model takes into account seven LVs in reflective measurement mode. Two LVs, namely "Quality" and "Image," have more than three MVs (7 and 5, respectively), while the tetrads for the remaining LVs were completed in line with the criteria described by Gudergan et al. (2008). Specifically, item CUSA1 was used to complete the set of MVs required to conduct CTA-PLS on "Expectation," "Loyalty," and "Value." The latter requires an additional MV (CUSA2) to complete the tetrad. The unique item "CUSCO" associated with the "Complaints" construct is included among the MVs that complete the "Satisfaction" LV's tetrad.

We report the model estimation in Fig. 6 provided by SmartPLS.

In line with the scope of the analysis of the empirical case studies, we extend our focus beyond the Bonferroni correction to include the Benjamini–Hochberg, Benjamini–Yekutieli, Hochberg, Holm, and Hommel corrections. From the outputs of the CTA-PLS, we derived the adjusted p values based on such corrections for each LV under investigation and each tetrad. Computations were carried out with R, and the results are presented in Table 1 to allow a practical comparison with the results in Gudergan et al. (2008, Table 5). Specifically, we consider each LV admitting at least two model-implied tetrads associated with as many hypotheses to be tested; for the sake of completeness, we also specify each LV's composite reliability ϱ_C and variance inflation factor (VIF) that inform about distinguished validity aspects for reflective and formative constructs, respectively. For each tetrad, we report its estimate from the original sample (residual value or *Res. Value* in Table 1, as it represents the deviation from the theoretical value 0 under the null hypothesis) and its bias (i.e., the difference between the bootstrap mean and the sample value).

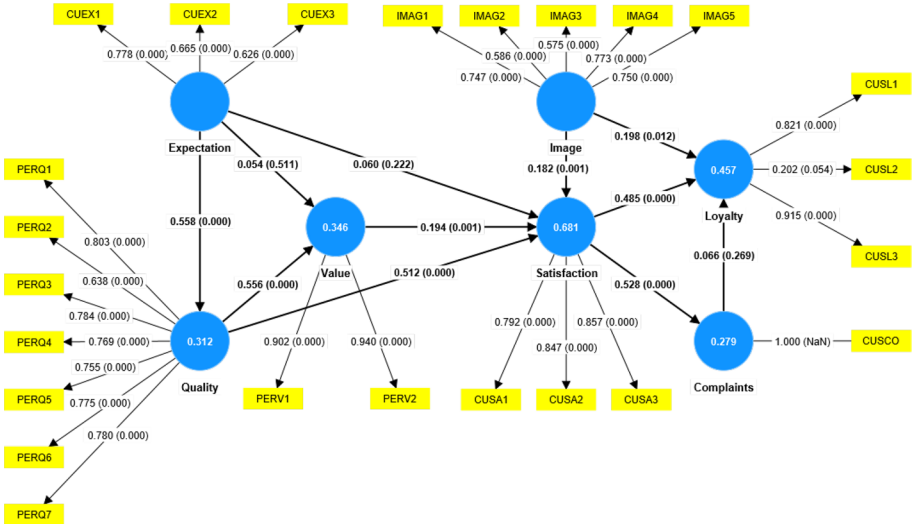


Fig. 6 The estimated ECSI model for mobile phone users (Gudergan et al., 2008) (5000 bootstrap resamples)

The corresponding t-statistics allow us to derive the tetrad-specific p values associated with each LV, and we report their values following six different adjustment methods under analysis. We underline those adjusted p values that are lower than the nominal significance level of 0.1.

We note that the composite reliability ϱ_C is greater than 0.7 for all the LVs, which agrees with the commonly accepted values for a reflective measurement model. Furthermore, the maximum VIF is lower than 3 for each construct, which supports the lack of potential multicollinearity issues (Hair et al., 2019). Similar properties also hold for "Value" ($\varrho_C = 0.918$, max VIF $= 1.962$), which is not presented since a single model-implied tetrad is generated through the criteria presented by Gudergan et al. (2008), so it does not enter our analysis of multiple hypothesis testing in CTA-PLS.

In this case study, disagreement between correction methods arises for two LVs. Indeed, from each correction except Benjamini–Yekutieli, one should reject the null hypothesis for the "Quality" construct due to the occurrence of a significant tetrad obtained from items PERQ1, PERQ2, PERQ3, and PERQ7 (Bonferroni-adjusted CI [0.039; 2.272]); on the other hand, the Benjamini–Yekutieli correction does not reject the hypothesis for this specific test (adjusted p value equal to 0.275). An analogous behaviour is observed for "Expectation," in particular for the model-implied tetrad obtained from CUEX1, CUEX2, CUSA1, and CUEX3, whose adjusted p value is 0.12 based on the Benjamini–Yekutieli correction and 0.08 for the others. If we assume that all the adjusted p values should be higher than the nominal significance level $\alpha = 0.1$, in line with the conjunctive form of H_0, then the Benjamini–Yekutieli correction does not entail sufficient evidence to reject the reflective measurement hypothesis for "Quality" and "Expectation," while H_0 should be rejected if one of the other corrections is selected. The behaviour of the Benjamini–Yekutieli method aligns with its performance in simulations (Sect. 3), and the agreement of the remaining methods suggests rejecting the null hypothesis for the "Quality" latent variable, as well as for "Expectation" and "Satisfaction."

Table 1 CTA-PLS tetrads and test results from the ECSI dataset ($n = 250$)

Res. value	Bias	t-stat.	CI (Bonferroni)	Adjusted p values					
				Bonferroni	Benjamini–Hochberg	Benjamini–Yekutieli	Hochberg	Holm	Hommel
Quality: 7 MVs, $\varrho_C = 0.905$, max VIF= 2.105									
0.172	− 0.005	0.537	[− 0.685; 1.039]	1.000	0.698	1.000	0.923	1.000	0.923
0.654	− 0.003	2.081	[− 0.188; 1.504]	0.524	0.175	0.568	0.449	0.449	0.375
0.651	− 0.009	2.174	[− 0.145; 1.465]	0.416	0.175	0.568	0.387	0.387	0.327
0.138	0.000	0.527	[− 0.565; 0.840]	1.000	0.698	1.000	0.923	1.000	0.923
1.139	− 0.016	2.746	[0.039; 2.272]	0.085	0.085	0.275	0.085	0.085	0.085
0.485	− 0.004	1.874	[− 0.207; 1.185]	0.854	0.213	0.694	0.671	0.671	0.549
0.217	− 0.004	0.844	[− 0.470; 0.912]	1.000	0.559	1.000	0.923	1.000	0.923
0.439	− 0.007	1.683	[− 0.256; 1.146]	1.000	0.232	0.755	0.895	0.925	0.740
− 0.347	0.001	0.977	[− 1.303; 0.607]	1.000	0.558	1.000	0.923	1.000	0.923
0.433	− 0.015	0.919	[− 0.820; 1.717]	1.000	0.558	1.000	0.923	1.000	0.923
− 0.438	0.005	1.648	[− 1.158; 0.273]	1.000	0.232	0.755	0.895	0.925	0.796
0.019	0.004	0.097	[− 0.516; 0.546]	1.000	0.923	1.000	0.923	1.000	0.923
− 0.286	0.004	0.978	[− 1.078; 0.497]	1.000	0.558	1.000	0.923	1.000	0.923
0.053	− 0.001	0.278	[− 0.456; 0.563]	1.000	0.841	1.000	0.923	1.000	0.923
Image: 5 MVs, $\varrho_C = 0.819$, max VIF= 1.510									
0.928	0.002	1.543	[− 0.473; 2.326]	0.615	0.307	0.702	0.492	0.492	0.466
1.152	− 0.006	1.832	[− 0.306; 2.622]	0.335	0.307	0.702	0.335	0.335	0.307
0.508	0.009	0.936	[− 0.763; 1.760]	1.000	0.436	0.997	0.576	0.722	0.576
0.204	− 0.010	0.560	[− 0.632; 1.060]	1.000	0.576	1.000	0.576	0.722	0.576
− 0.390	− 0.002	1.173	[− 1.162; 0.386]	1.000	0.401	0.916	0.576	0.722	0.524
Expectation: 3 MVs, $\varrho_C = 0.733$, max VIF= 1.161									
0.217	− 0.011	0.672	[− 0.406; 0.863]	1.000	0.502	0.753	0.502	0.502	0.502
0.548	− 0.009	2.055	[0.034; 1.081]	0.080	0.080	0.120	0.080	0.080	0.080

Table 1 (continued)

Res. value	Bias	t-stat.	CI (Bonferroni)	Adjusted p values					
				Bonferroni	Benjamini–Hochberg	Benjamini–Yekutieli	Hochberg	Holm	Hommel
Loyalty: 3 MVs, $\varrho_C = 0.722$, max VIF= 1.427									
0.230	0.006	0.344	[− 1.088; 1.537]	1.000	0.731	1.000	0.731	1.000	0.731
0.217	− 0.009	0.373	[− 0.914; 1.366]	1.000	0.731	1.000	0.731	1.000	0.731
Satisfaction: 3 MVs, $\varrho_C = 0.871$, max VIF= 1.762									
0.784	− 0.003	2.279	[0.112; 1.462]	0.045	0.044	0.066	0.044	0.045	0.044
0.837	− 0.010	2.015	[0.032; 1.660]	0.088	0.044	0.066	0.044	0.045	0.044

Adjusted p values less than or equal to 0.1 are underlined. A double line separates tetrads derived from LVs with at least 4 MVs from other model-implied tetrads

4.2 The BCP Dataset and Model

We consider a second empirical application by examining the data shared and analysed by Damberg (2023). The author discusses the main aspect of the model explored by Damberg et al. (2022), which concentrates on consumer perceptions in the banking sector, specifically cooperative banks. The model pays special attention to relations between customer satisfaction, relational trust, and loyalty.

Compared with the previous models, the present one is more complex in terms of the number of LVs with more than three MVs and the tetrads per LV. In particular, five of the nine LVs have at least four MVs; they are denoted as "*Perceived Quality*" ("Qual," 6 MVs), "*Perceived Performance*" ("Perf," 5 MVs), "*Perceived Corporate Social Responsibility*" ("Csor," 5 MVs), "*Perceived Attractiveness*" ("Attr," 4 MVs), and "*Relational Trust* ("Trust," 4 MVs). All these LVs are measured in formative mode, except for the reflectively measured "Trust." The remaining variables with less than four MVs are "*Perceived Competence*" ("Comp," 3 MVs), *Perceived Likeability* ("Like," 2 MVs), "*Customer Satisfaction*" ("Sat," 3 MVs), and "*Customer Loyalty*" ("Loy," 3 MVs); all of them are measured in reflective mode. The sample presented in the work of Damberg (2023) is composed of 675 responses, and all the items are measured on a 7-point scale.

As for the previous case study, in Fig. 7, we report estimates and associated p values derived for this model from PLS-SEM bootstrapping.

In Table 2, we present the summary of the CTA-PLS test and the p values associated with each tetrad, aligning with the information reported in Table 1 to enhance the comparability of evidence extracted from the computation. Here, we also specify the inclusion of model-implied tetrads for such LVs with three MVs by exploiting the approach suggested by Gudergan et al. (2008). Specifically, we use the Fornell-Larcker criterion: successors are examined first, if they exist, and predecessors are considered otherwise. Then, we choose those MVs from the latter construct that achieve maximal cross-loadings with the former (Cheah et al., 2019; Puche-Regaliza et al., 2021). While this criterion may be of interest in the present analysis, we stress that recent work suggests relying on LVs with at

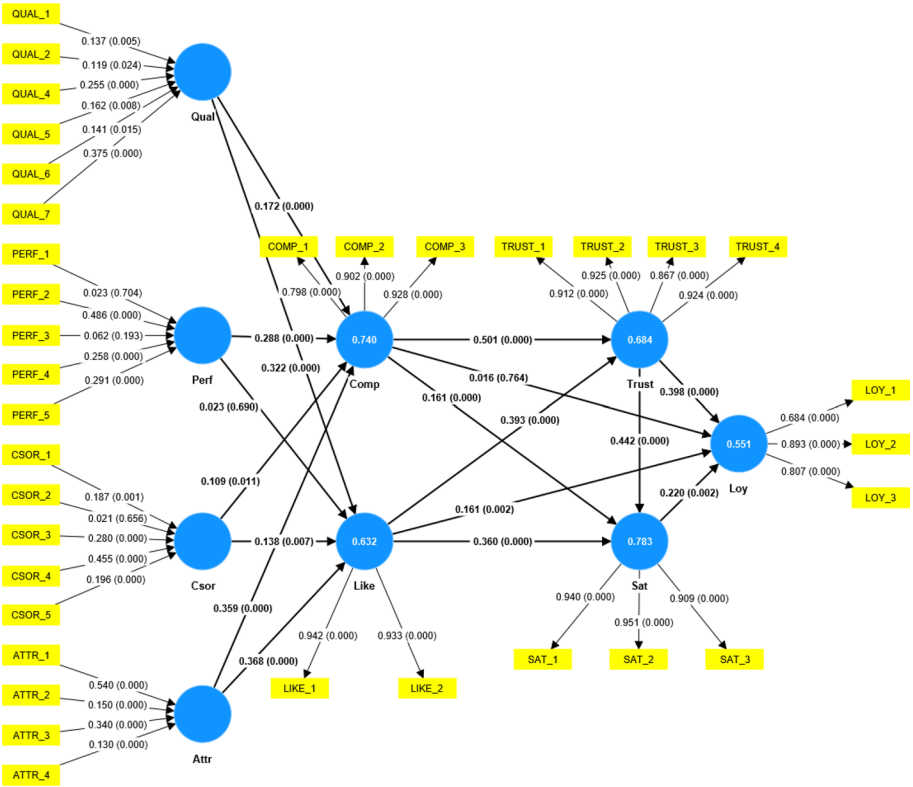


Fig. 7 The estimated BCP model (Damberg et al., 2022) (5000 bootstrap resamples)

least four MVs without extending the analysis to model-implied tetrads (Hair et al., 2017; Teeluckdharry et al., 2022). Therefore, we complete the tetrads for LVs with less than four MVs by including TRUST_4 among the MVs of both "Comp" and "Loy," while LOY_2 and the pair (SAT_2, SAT_3) complete "Sat" and "Like," respectively.

All the LVs measured in reflective mode show high composite reliability. The remaining LVs are measured in the formative mode, which does not include a measurement error underlying the construction of the composite reliability (Diamantopoulos, 2006), and they show acceptable values of VIF (below 5). The analysis reveals that the CTA-PLS does not reject the reflective measurement hypothesis for "Attr" and "Csor," which, however, are measured in a formative mode in the original model. The magnitude of the p values obtained from (5) makes all the adjustment corrections consistent. Dually, the "Qual" and "Trust" MVs show significant deviations from the outputs expected from reflective measurement modes. This supports the formative measurement mode for "Qual," as in Damberg (2022); Damberg et al. (2023), but also for "Trust," even if it is not an antecedent.

This model includes a single inconsistency among adjustment corrections, which is associated with the tetrad $\tau_{1352}^{\text{Perf}}$. In agreement with the results discussed above, Benjamini–Hochberg tends to reject the null hypothesis for this tetrad, while Benjamini–Yekutieli and Bonferroni do not provide enough evidence for the rejection of H_0. In particular, if we reject the reflective measurement mode for "Perf" based on the other

Table 2 Adjusted *p* values for each LV and tetrad in the BCP model based on the different corrections

Res. value	Bias	t-stat	CI (Bonferroni)	Adjusted *p* values					
				Bonferroni	Benjamini–Hochberg	Benjamini–Yekutieli	Hochberg	Holm	Hommel
Qual: 6 MVs, max VIF= 3.878									
0.036	− 0.001	0.572	[− 0.124; 0.197]	1.000	0.729	1.000	0.816	1.000	0.816
0.200	− 0.001	4.156	[0.079; 0.324]	0.000	0.000	0.000	0.000	0.000	0.000
− 0.022	0.000	0.304	[− 0.203; 0.160]	1.000	0.816	1.000	0.816	1.000	0.816
0.227	− 0.001	3.873	[0.079; 0.377]	0.001	0.000	0.001	0.001	0.001	0.001
− 0.014	0.000	0.233	[− 0.167; 0.139]	1.000	0.816	1.000	0.816	1.000	0.816
0.229	− 0.001	3.282	[0.053; 0.407]	0.009	0.002	0.005	0.005	0.005	0.004
0.400	− 0.001	6.098	[0.235; 0.568]	0.000	0.000	0.000	0.000	0.000	0.000
0.174	0.000	3.226	[0.037; 0.311]	0.011	0.002	0.005	0.005	0.005	0.005
0.284	− 0.001	3.789	[0.095; 0.475]	0.001	0.000	0.001	0.001	0.001	0.001
Csor: 5 MVs, max VIF= 3.029									
0.079	− 0.001	1.767	[− 0.024; 0.184]	0.386	0.129	0.294	0.232	0.232	0.232
− 0.076	− 0.001	1.347	[− 0.207; 0.057]	0.890	0.223	0.508	0.256	0.356	0.256
0.096	− 0.001	1.920	[− 0.019; 0.213]	0.275	0.129	0.294	0.220	0.220	0.165
− 0.108	0.000	2.206	[− 0.221; 0.006]	0.137	0.129	0.294	0.137	0.137	0.129
0.061	− 0.001	1.136	[− 0.063; 0.186]	1.000	0.256	0.585	0.256	0.356	0.256
Perf: 5 MVs, max VIF= 3.625									
0.037	0.000	0.956	[− 0.053; 0.126]	1.000	0.424	0.968	0.546	0.678	0.546
0.141	0.000	4.423	[0.067; 0.216]	0.000	0.000	0.000	0.000	0.000	0.000
− 0.026	− 0.001	0.604	[− 0.123; 0.074]	1.000	0.546	1.000	0.546	0.678	0.546
0.090	0.000	2.154	[− 0.007; 0.188]	0.157	0.052	0.119	0.094	0.094	0.094
0.231	0.000	5.143	[0.127; 0.336]	0.000	0.000	0.000	0.000	0.000	0.000
Attr: 4 MVs, max VIF= 2.580									
0.021	0.000	0.283	[− 0.126; 0.168]	1.000	0.960	1.000	0.960	1.000	0.960
0.003	0.000	0.051	[− 0.122; 0.128]	1.000	0.960	1.000	0.960	1.000	0.960

Table 2 (continued)

Res. value	Bias	t-stat	CI (Bonferroni)	Adjusted p values					
				Bonferroni	Benjamini–Hochberg	Benjamini–Yekutieli	Hochberg	Holm	Hommel
Trust: 4 MVs, $\varrho_C = 0.949$, max VIF= 4.912									
0.512	− 0.002	7.589	[0.382; 0.646]	0.000	0.000	0.000	0.000	0.000	0.000
0.448	− 0.001	6.362	[0.311; 0.587]	0.000	0.000	0.000	0.000	0.000	0.000
Comp: 3 MVs, $\varrho_C = 0.909$, max VIF= 2.864									
0.034	− 0.002	0.755	[− 0.052; 0.123]	0.900	0.450	0.675	0.450	0.881	0.450
0.041	− 0.001	0.772	[− 0.061; 0.145]	0.881	0.450	0.675	0.450	0.881	0.450
Loy: 3 MVs, $\varrho_C = 0.840$, max VIF= 1.896									
− 0.054	0.000	0.574	[− 0.238; 0.130]	1.000	0.566	0.849	0.566	0.566	0.566
− 0.401	0.002	4.046	[− 0.597; -0.209]	0.000	0.000	0.000	0.000	0.000	0.000
Sat: 3 MVs, $\varrho_C = 0.953$, max VIF=4.903									
0.151	− 0.001	2.206	[0.018; 0.287]	0.055	0.055	0.082	0.055	0.055	0.055
0.112	− 0.001	1.524	[− 0.031; 0.258]	0.255	0.128	0.192	0.128	0.128	0.128

Adjusted p values less than or equal to 0.1 are underlined. A double line separates tetrads derived from LVs with at least 4 MVs from other model-implied tetrads

tetrads (e.g., $\tau^{\text{Perf}}_{1345}$, which is highly significant for all the corrections), Bonferroni and Benjamini–Yekutieli do not recognise the additional evidence given by the tetrad $\tau^{\text{Perf}}_{1352}$. For the BCP model, the tetrads where adjustment methods disagree do not compromise the overall agreement at the decision level. Specifically, we can consistently reject the reflective model hypothesis for all the constructs displayed in Table 2 with the exception of "Csor," "Attr," and "Comp."

The following steps summarise the results from simulations and empirical applications, supporting the selection of measurement models in various scenarios:

- Evaluate the proposed model's validity based on the specific measurement mode (composite reliability, AVE, VIF, *etc*.).
- Specify the criteria for identifying model-implied tetrads, as it was done for the BCP model. In particular, choose whether to include model-implied tetrads in the testing procedure (Gudergan et al., 2008; Cheah et al., 2019; Puche-Regaliza et al., 2021) or not (Hair et al., 2017; Teeluckdharry et al., 2022).
- Use resampling or bootstrap procedures to compute p values and adjust them using multiple criteria based on different approaches (e.g., FWER, FDR).

- If all adjustment methods lead to the same results, apply Bonferroni to decide whether to reject H_0, as it aligns with the conjunctive logic of the null hypothesis in CTA-PLS, where a reflective model implies that all tetrads simultaneously vanish.
- If the adjustment methods lead to different results:
  - favour Bonferroni or other FWER methods if the sample size is large ($n > 400$), there is theoretical evidence of a reflective model, or there are few MVs (low measurement model complexity), as a stronger control of Type I errors may outweigh a limited loss of power in these cases.
  - Favour Benjamini–Hochberg if the model is complex in terms of MVs, the sample size is small, or there are not strong theoretical motivations for reflective measurements. Additionally, consider assessing consistency across adjustment methods or multiple datasets using adapted resampling procedures.

The proposed selection process, depicted in Fig. 8, extends CTA-PLS to support psychological and social research when the distinction between different measurement modes is vague or unsharp. Remarkably, some of the conditions that favour FDR over FWER approaches, such as small sample size, complex models, or limited theoretical evidence excluding formative indicators, also motivate the choice of PLS-SEM; in this sense, the proposed approach complements and refines existing guidelines for proper use of PLS-SEM in research (Hair et al., 2019, pp. 4–5), especially when powerful adjustment methods are needed to uncover potential discoveries in line with the causal-predictive rationale underlying PLS-SEM.

5 Concluding Remarks

This work was motivated by the need for a deeper investigation of the effects of different correction methods for multiple hypothesis testing in measurement mode specification. Such effects were explored in detail with the aim of supporting researchers and practitioners of SEM by providing evidence obtained from both simulated and empirical data. The combination of these two approaches allows encompassing a variety of scenarios characterised by sample size and model complexity (with special regard to the number of MVs generating tetrads) while connecting the results from simulations to real-world case studies. A data generation process was designed for reflective and formative measurement models by testing the actual significance level and actual power in CTA-PLS for different sample sizes. In addition to the classical Bonferroni correction, which is implemented in the well-known SmartPLS software, the p value of the multiple tests has been adjusted with the Benjamini–Hochberg and other correction approaches.

For what concerns the reflective simulations, it emerged that Bonferroni commits a slightly (but significantly) lower probability of Type I error than Benjamini–Hochberg; in addition, by increasing the number of manifest variables in the measurement model, the probability of Type I error of the CTA-PLS decreases for small sample sizes. Indeed, from medium to high sample sizes, the actual significance level seems to be independent of the number of manifest variables. Concerning the formative simulations, Benjamini–Hochberg seems to be more powerful than Bonferroni. Another macro-evidence is that the actual test power grows by increasing the sample size for all the measurement models and adjustments. In addition, by increasing the number of manifest variables in the measurement model, the gap of actual power between small and medium-large sample sizes increases;

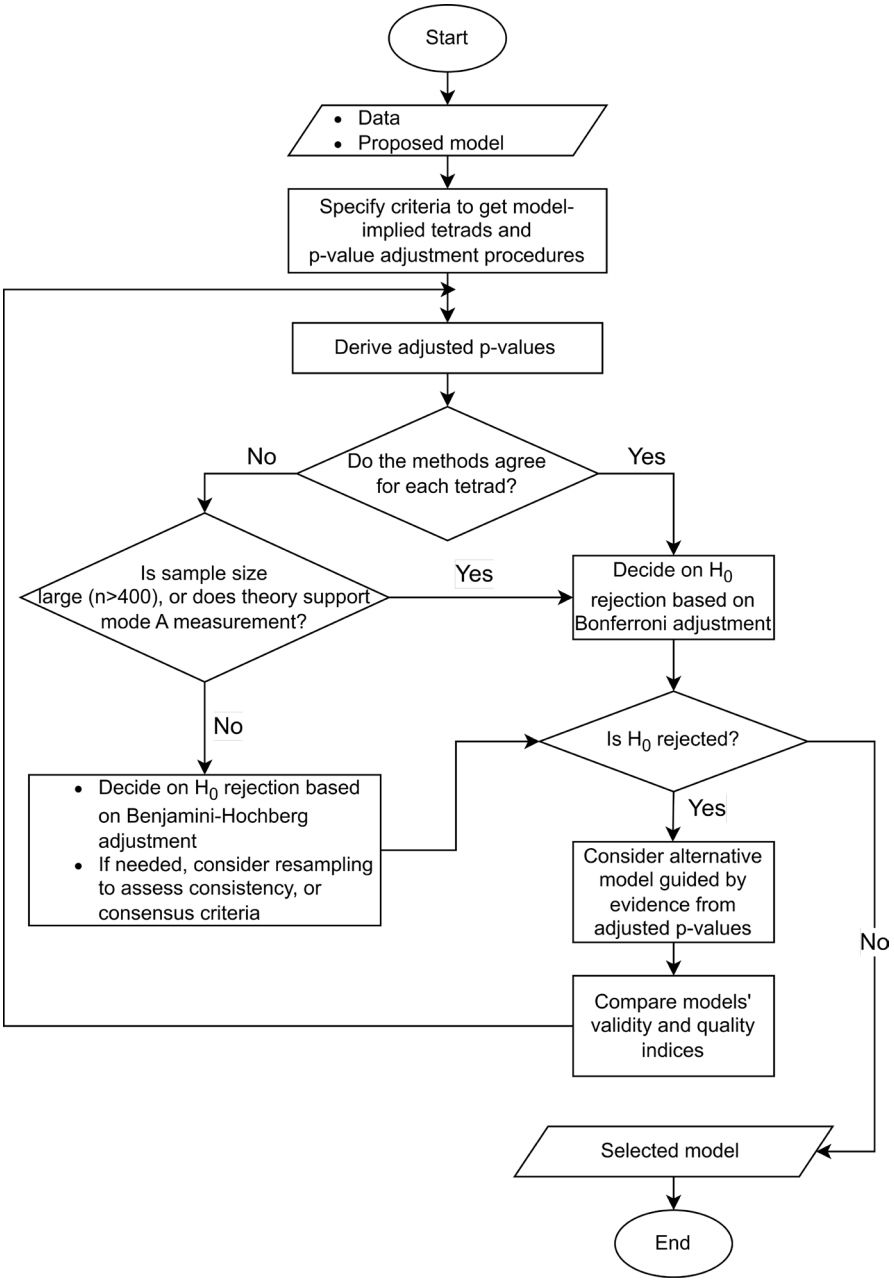


Fig. 8 Diagrammatic representation of the measurement model decision process based on CTA-PLS and multiple adjustment methods

the curve slope (for both adjustment approaches) grows faster by increasing the number of MVs.

Then, two empirical applications were treated based on three main factors, namely, data accessibility, the association with a measurement model with a suitable number of indicators to conduct CTA-PLS, and the dimensionality of the dataset. The first dataset used was the European Customers Satisfaction Index (ECSI) for mobile phone users (2005), whose analysis was in continuity with Gudergan et al. (2008) ($n = 250$). In this case study, disagreement between correction methods arose for two LVs; in fact, Benjamini–Yekutieli correction did not entail sufficient evidence to reject the reflective measurement hypothesis, despite the other methods. The second empirical application examined the data shared and analysed by Damberg (2023) (more complex in terms of the number of LVs with more than three MVs, $n = 675$). In this case, CTA-PLS did not reject the reflective measurement hypothesis for two LVs, which were measured in a formative mode in the original model. It included a single inconsistency among adjustment corrections, which was associated with just one tetrad; Benjamini–Hochberg rejected the null hypothesis for this one, while Benjamini–Yekutieli and Bonferroni did not provide enough evidence for its rejection. Additional analyses regarding the case studies made us focus on the agreement between different correction methods in terms of rejection of H_0. When the null hypothesis is rejected, we can obtain multiple subsamples from the original dataset to evaluate the frequency of null hypothesis rejection from different methods, which allows us to extract more information on the sensitivity of the test in empirical scenarios. This approach will be explored in more detail in future work.

Overall, we remark that the Bonferroni correction is too conservative and has lower power, especially with small sample sizes and a high number of indicators, with respect to the Benjamini–Hochberg correction. Other ones are similar to or equal to Bonferroni or have bad performances (the Benjamini–Yekutieli correction). In particular, if the researcher's primary goal is to minimize the type I error, the conservative Bonferroni method is recommended. However, when maximizing test power is the focus, especially with small sample sizes or complex structures, the Benjamini–Hochberg correction is preferable: while it offers significantly higher power, it still holds to the control of type I error, making it a generally more balanced and practical choice compared to the highly conservative Bonferroni approach.

For future developments, it could be interesting to deepen this work by considering more complex measurement models, e.g. by increasing the number of MVs in the measurement model (and, hence, the number of multiple tests). This extension could be relevant to stress the differences in the multiple correction adjustments performance. Another interesting step forward could be taking into account hierarchical models with structure higher than a second-order (Cefis & Carpita, 2024), for generalize CTA-PLS to higher-order constructs.

Appendix: Random Correlation Matrix Generation

Correlation matrices with the same composite reliability for the alternative hypothesis (formative measurement model) are generated using the approach of Joe (2006) with the function *rcorrmatrix* in the *R* package *clusterGeneration* version 1.3.8. With this approach each off-diagonal correlation has a non-standard $Beta(\alpha, \alpha)$ distribution on $(-1; +1)$, where $\alpha = d + (v - 2)/2$ with v is the number of variables and the parameter $d > 0$; when $d = 1$ the

random matrix is uniform over the space of positive definite correlation matrices. We have adapted this general procedure to our case, as explained in the following. As the non-standard $Beta(\alpha, \beta)$ random variable X on (inf, sup) is a linear transformation of the standard $Beta(\alpha, \beta)$ random variable Z on (0, 1):

$$X = (sup - inf) \cdot Z + inf \tag{7}$$

we obtain

$$E(X) = (sup - inf) \cdot \alpha/(\alpha + \beta) + inf \tag{8}$$

and

$$Var(X) = (sup - inf)^2 \cdot \alpha \cdot \beta/[(\alpha + \beta + 1) \cdot (\alpha + \beta)^2]. \tag{9}$$

For the function *rcorrmatrix* $inf = -1$, $sup = +1$ and $\alpha = \beta$, so that in this special case $X = 2 \cdot Z - 1$ and therefore:

$$E(X) = 0 \text{ and } Var(X) = 1/(2 \cdot \alpha + 1). \tag{10}$$

The model used by Gudergan et al. (2008) has 3 LVs each with 5 MVs, and the 3 correlation matrices have off-diagonal minimum absolute correlation of 0.1 [to avoid correlation close to zero in the measurement model; Bollen and Ting (2000)], with off-diagonal elements that have $min = -0.15$, $max = 0.72$, $mean = 0.22$ and $St.dev = 0.24$. To simulate an analogous situation, we created the random correlation matrix with generic off-diagonal element obtained applying the linear transformation (A1) to the standard Beta random variable Z with $inf = 0$ and $sup = 0.5$: by substituting those values to (8) and (9) we obtain $E(X) = 0.25$ and $Var(X) = 0.25/[4 \cdot (2 \cdot \alpha + 1)]$. Finally, to avoid correlations close to zero in the measurement model, we added 0.1 to correlations lower than 0.1 (Gudergan et al., 2008). As summary, for each replication we generated a random correlation matrix with $d = 0.1$, $inf = 0$, $sup = 0.5$ and v from 4 to 8, with the constraint that the composite reliability index is $\rho_c = 0.7$ (Gudergan et al., 2008).

Funding Open access funding provided by Università degli Studi di Brescia within the CRUI-CARE Agreement. M.A. acknowledges financial support from the Project Small Area Estimation for data Quality in Health - SAEQHealth (Bando a cascata Programma PE GRINS - GRINS - Growing Resilient, Inclusive and Sustainable (cod. PE0000018).

Data Availability The data have not been uploaded to a public repository; however, the authors are willing to share them upon request.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Ethics statement All the research was conducted in accordance with ethical guidelines and approved protocols, ensuring compliance with relevant standards and regulations.

References

Benjamini, Y., Heller, R., & Yekutieli, D. (2009). Selective inference in complex research. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 367*(1906), 4255–4271. https://doi.org/10.1098/rsta.2009.0127

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics, 29*(4), 1165–1188. https://doi.org/10.1214/aos/1013699998

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53*(1), 605–634. https://doi.org/10.1146/annurev.psych.53.100901.135239

Bollen, K. A., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110*(2), 305–314. https://doi.org/10.1037/0033-2909.110.2.305

Bollen, K. A., Lennox, R. D., & Dahly, D. L. (2009). Practical application of the vanishing tetrad test for causal indicator measurement models: An example from healthrelated quality of life. *Statistics in Medicine, 28*(10), 1524–1536. https://doi.org/10.1002/sim.3560

Bollen, K. A., & Ting, K.-F. (1993). Confirmatory tetrad analysis. *Sociological Methodology, 23*, 147–175. https://doi.org/10.2307/271009

Bollen, K. A., & Ting, K.-F. (1998). Bootstrapping a test statistic for vanishing tetrads. *Sociological Methods & Research, 27*(1), 77–102. https://doi.org/10.1177/0049124198027001002

Bollen, K. A., & Ting, K.-F. (2000). A tetrad test for causal indicators. *Psychological Methods, 5*(1), 3–22. https://doi.org/10.1037/1082-989x.5.1.3

Cavicchia, C., & Vichi, M. (2021). Statistical model-based composite indicators for tracking coherent policy conclusions. *Social Indicators Research, 156*(2–3), 449–479. https://doi.org/10.1007/s11205-020-02318-7

Cefis, M., & Carpita, M. (2022). The higher-order PLS-SEM confirmatory approach for composite indicators of football performance quality. *Computational Statistics, 39*(1), 93–116. https://doi.org/10.1007/s00180-022-01295-4

Cefis, M., & Carpita, M. (2024). On the CTA-PLS test for hierarchical models: An application to the football player's performance. *Computational Statistics*. https://doi.org/10.1007/s00180-024-01566-2

Chang, W., Franke, G. R., & Lee, N. (2016). Comparing reflective and formative measures: New insights from relevant simulations. *Journal of Business Research, 69*(8), 3177–3185. https://doi.org/10.1016/j.jbusres.2015.12.006

Cheah, J.-H., Roldán, J. L., Ciavolino, E., Ting, H., & Ramayah, T. (2021). Sampling weight adjustments in partial least squares structural equation modeling: guidelines and illustrations. *Total Quality Management & Business Excellence, 32*(13–14), 1594–1613. https://doi.org/10.1080/14783363.2020.1754125

Cheah, J.-H., Ting, H., Ramayah, T., Memon, M. A., Cham, T.-H., & Ciavolino, E. (2019). A comparison of five reflective-formative estimation approaches: Reconsideration and recommendations for tourism research. *Quality & Quantity, 53*(3), 1421–1458. https://doi.org/10.1007/s11135-018-0821-7

Ciavolino, E., Angelelli, M., Sternativo, G. A., De Carlo, E., Catalano, A. A., & Ingusci, E. (2024). A higher-order job crafting mediation model with PLS-SEM: Relationship between organizational identification and communication satisfaction. *Soft Computing*. https://doi.org/10.1007/s00500-024-09667-2

Ciavolino, E., Aria, M., Cheah, J.-H., & Roldán, J. L. (2022). A tale of PLS structural equation modelling: Episode I-A bibliometrix citation analysis. *Social Indicators Research, 164*(3), 1323–1348. https://doi.org/10.1007/s11205-022-02994-7

Ciavolino, E., Carpita, M., & Nitti, M. (2015). High-order PLS path model with qualitative external information. *Quality & Quantity, 49*(4), 1609–1620. https://doi.org/10.1007/s11135-014-0068-x

Ciavolino, E., Ferrante, L., Sternativo, G. A., Cheah, J.-H., Rollo, S., Marinaci, T., & Venuleo, C. (2022). A confirmatory composite analysis for the Italian validation of the interactions anxiousness scale: A higher-order version. *Behaviormetrika, 49*(1), 23–46. https://doi.org/10.1007/s41237-021-00151-x

Coltman, T., Devinney, T. M., Midgley, D. F., & Venaik, S. (2008). Formative versus reflective measurement models: Two applications of formative measurement. *Journal of Business Research, 61*(12), 1250–1262. https://doi.org/10.1016/j.jbusres.2008.01.013

Damberg, S. (2023). Advanced PLS-SEM models for bank customer relationship management using survey data. *Data in Brief, 48*, 109187. https://doi.org/10.1016/j.dib.2023.109187

Damberg, S., Schwaiger, M., & Ringle, C. M. (2022). What's important for relationship management? The mediating roles of relational trust and satisfaction for loyalty of cooperative banks' customers. *Journal of Marketing Analytics, 10*(1), 3–18. https://doi.org/10.1057/s41270-021-00147-2

Danks, N. P., Sharma, P. N., & Sarstedt, M. (2020). Model selection uncertainty and multimodel inference in partial least squares structural equation modeling (PLSSEM). *Journal of Business Research, 113*, 13–24. https://doi.org/10.1016/j.jbusres.2020.03.019

Diamantopoulos, A. (2006). The error term in formative measurement models: interpretation and modeling implications. *Journal of Modelling in Management, 1*(1), 7–17. https://doi.org/10.1108/17465660610667775

Diamantopoulos, A., Riefler, P., & Roth, K. P. (2008). Advancing formative measurement models. *Journal of Business Research, 61*(12), 1203–1218. https://doi.org/10.1016/j.jbusres.2008.01.009

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research, 38*(2), 269–277. https://doi.org/10.1509/jmkr.38.2.269.18845

Dolce, P., Davino, C., & Vistocco, D. (2022). Quantile composite-based path modeling: Algorithms, properties and applications. *Advances in Data Analysis and Classification, 16*(4), 909–949. https://doi.org/10.1007/s11634-021-00469-0

Gudergan, S. P., Ringle, C. M., Wende, S., & Will, A. (2008). Confirmatory tetrad analysis in PLS path modeling. *Journal of Business Research, 61*(12), 1238–1249. https://doi.org/10.1016/j.jbusres.2008.01.012

Hair, J. F., Howard, M. C., & Nitzl, C. (2020). Assessing measurement model quality in PLS-SEM using confirmatory composite analysis. *Journal of Business Research, 109*, 101–110. https://doi.org/10.1016/j.jbusres.2019.11.069

Hair, J. F., Risher, J. J., Sarstedt, M., & Ringle, C. M. (2019). When to use and how to report the results of PLS-SEM. *European Business Review, 31*(1), 2–24. https://doi.org/10.1108/ebr-11-2018-0203

Hair, J. F., Sarstedt, M., Ringle, C. M., & Gudergan, S. P. (2017). *Advanced issues in partial least squares structural equation modeling*. Thousand Oaks, CA: SAGE Publications.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*(4), 800–802. https://doi.org/10.2307/2336325

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika, 75*(2), 383–386. https://doi.org/10.2307/2336190

Ingusci, E., Angelelli, M., Sternativo, G. A., Catalano, A. A., De Carlo, E., Cortese, C. G., & Ciavolino, E. (2024). A higher-order life crafting scale validation using PLS-CCA: The Italian version. *Behaviormetrika, 51*(1), 359–387. https://doi.org/10.1007/s41237-023-00209-y

Ingusci, E., Signore, F., De Carlo, E., & Angelelli, M. (2023). Human resources management practices and job satisfaction: The moderating role of seeking challenges. A longitudinal study through PLS-SEM. *Electronic Journal of Applied Statistical Analysis, 16*(1), 25–49. https://doi.org/10.1285/i20705948v16n1p25

Joe, H. (2006). Generating random correlation matrices based on partial correlations. *Journal of Multivariate Analysis, 97*(10), 2177–2189. https://doi.org/10.1016/j.jmva.2005.05.010

Jöreskog, K. G. (1973). *Analysis of covariance structures Multivariate Analysis-III* (pp. 263–285). Amsterdam, Netherlands: Elsevier.

Mendes, T., Braga, V., Silva, C., & Braga, A. (2023). The speed of internationalization in regionally clustered family firms: A deeper understanding of innovation activities and cluster affiliation. *Review of Regional Research*. https://doi.org/10.1007/s10037-023-00182-9

Nunnally, J. C. (1994). *Psychometric theory 3E*. New York, NY: Tata McGraw-Hill education.

Ongena, G. (2023). Data literacy for improving governmental performance: A competence-based approach and multidimensional operationalization. *Digital Business, 3*(1), 100050. https://doi.org/10.1016/j.digbus.2022.100050

Puche-Regaliza, J. C., Porras-Alfonso, S., Jiménez, A., Aparicio-Castillo, S., & Arranz- Val, P. (2021). Exploring determinants of public satisfaction with urban solid waste collection services quality. *Environment, Development and Sustainability, 23*, 9927–9948. https://doi.org/10.1007/s10668-020-01040-1

Rajala, R., & Westerlund, M. (2010). Antecedents to consumers' acceptance of mobile advertisements-A hierarchical construct PLS structural equation model. In *2010 43rd Hawaii International Conference on System Sciences* (pp. 1–10).

Ringle, C., Da Silva, D., & Bido, D. (2015). Structural equation modeling with the SmartPLS. *Brazilian Journal of Marketing, 13*(2), 56–73. https://doi.org/10.5585/remark.v13i2.2717

Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O., & Gudergan, S. P. (2016). Estimation issues with PLS and CBSEM: Where the bias lies! *Journal of Business Research, 69*(10), 3998–4010. https://doi.org/10.1016/j.jbusres.2016.06.007

Sarstedt, M., Ringle, C. M., & Hair, J. F. (2021). Partial least squares structural equation modeling. *Handbook of market research* (pp. 587–632). Heidelberg, Germany: Springer.

Schamberger, T. (2023). Conducting Monte Carlo simulations with PLS-PM and other variance-based estimators for structural equation models: A tutorial using the R package cSEM. *Industrial Management & Data Systems, 123*(6), 1789–1813. https://doi.org/10.1108/imds-07-2022-0418

Schlittgen, R., Sarstedt, M., & Ringle, C. M. (2020). Data generation for compositebased structural equation modeling methods. *Advances in Data Analysis and Classification, 14*(4), 747–757. https://doi.org/10.1007/s11634-020-00396-6

Simonetto, A. (2012). Formative and reflective models: State of the art. *Electronic Journal of Applied Statistical Analysis, 5*(3), 452–457. https://doi.org/10.1285/i20705948v5n3p452

Tabet, S. M., Lambie, G. W., Jahani, S., & Rasoolimanesh, S. M. (2020). An analysis of the world health organization disability assessment schedule 2.0 measurement model using partial least squares-structural equation modeling. *Assessment, 27*(8), 1731–1747. https://doi.org/10.1177/1073191119834653

Tabet, S. M., Lambie, G. W., Jahani, S., & Rasoolimanesh, S. M. (2020). The factor structure of outcome questionnaire–45.2 scores using confirmatory tetrad analysis-partial least squares. *Journal of Psychoeducational Assessment, 38*(3), 350–368. https://doi.org/10.1177/0734282919842035

Teeluckdharry, N. B., Teeroovengadum, V., & Seebaluck, A. K. (2022). A roadmap for the application of PLS-SEM and IPMA for effective service quality improvements. *The TQM Journal, 36*(5), 1300–1345. https://doi.org/10.1108/tqm-11-2021-0340

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini–Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27*(1), 77–83. https://doi.org/10.3102/10769986027001077

Wold, H. (1985). Partial least squares. In S. Kotz & N. Johnson (Eds.), *Encyclopedia of statistical sciences* (Vol. 6, pp. 581–591). New York, NY: Wiley.