Received: 15 October 2025 | Revised: 28 November 2025 | Accepted: 5 February 2026 | Published online: 3 March 2026

RESEARCH ARTICLE

International Journal of Changes in Education
2026, Vol. 00(00) 1–9
DOI: 10.47852/bonviewIJCE62027911

BON VIEW PUBLISHING

Exploring Methodological Challenges in Educational Research: A Case Reflection on Response Surface Analysis

Kim Sorjonen[1,*] and Rebeka Weidmann[2]

[1] *Department of Psychology, Stockholm University, Sweden*

[2] *University of Basel, Switzerland*

Abstract: The growing availability of open-access analysis scripts has facilitated the use of advanced statistical methods such as response surface analysis (RSA). However, accessibility alone does not ensure appropriate application. In this commentary, we consider a published study as one illustrative case study, showing how the use of shared scripts—even from expert sources—may lead to fragile conclusions if methodological assumptions are overlooked. Through a detailed reconstruction of the modeling process, we identify mismatches between analytic choices and theoretical aims, including issues in model specification, data treatment, and interpretation. A small simulation further illustrates how routine preprocessing decisions, such as handling missing data or centering predictors, can meaningfully alter the fitted surface, underscoring the practical consequences of these choices. We offer practical suggestions to help researchers identify common pitfalls in the use of RSA. We also place these observations in the broader context of open science, highlighting how transparency in analysis must be accompanied by careful scrutiny of modeling assumptions. This reflection aims to support more enlightened analytic practices and encourage cautious engagement with complex statistical techniques. In doing so, we argue that the benefits of open-access tools can be fully realized if paired with clear methodological awareness and sound theoretical reasoning.

Keywords: education methods, response surface analysis, case study, reproducibility, open science

1. Introduction

> "*With great power there must*
> *also come—great responsibility!"*
> Lee, S., & Ditko, S. (1962)[1].
> *Amazing Fantasy (#15)*.
> Marvel Comics.

The development of open-access data and scripts has contributed to major advances in the dissemination of advanced modeling techniques in the social sciences, including education. This progress reflects growing calls for data and code sharing to improve reproducibility, particularly in disciplines facing concerns about replicability (e.g., psychology) [1, 2]. It is further supported by platforms for open dissemination (e.g., Open Science Framework) and statistical software environments such as R [3]. However, while advanced techniques can now be implemented more easily using openly shared scripts, their correct application and interpretation still require a solid understanding of the statistical concepts involved. Indeed, with great power of statistical analysis comes great responsibility of statistical knowledge.

A recent study by Leroy et al. [4] on stereotype effects at school offers a useful case for illustrating challenges that can arise when complex methods are applied without sufficient attention to underlying statistical assumptions. The study used response surface analysis (RSA), a technique combining polynomial regression and graphical interpretation to examine how outcomes vary across combinations of two predictors. While RSA has gained increasing attention in recent years [5–7], it also demands careful attention to modeling assumptions and interpretation. In this case, the analysis relied on R scripts provided by RSA developers [5], yet several methodological challenges remained—suggesting that open scripts alone may not be sufficient in the absence of deeper statistical understanding. The study's publication in a high-impact journal *Learning and Instruction* further underscores how such issues may persist even through peer review. Importantly, the challenges discussed here are not isolated but can be found—albeit more scattered—in other recent studies using RSA [8–10], pointing to broader issues in the applied use of this technique.

The purpose of this article is to examine these challenges through a detailed reconstruction of one published RSA. This case was selected because the availability of open materials and

***Corresponding author:** Kim Sorjonen, Department of Psychology, Stockholm University, Sweden. Email: ksorjonen.psych@gmail.com; kim.sorjonen@su.se

[1] Steve Ditko and Stan Lee, "Amazing Fantasy (#15)", Internet Archive, 1962. https://archive.org/details/Amazing_Fantasy_vol1_15_201607/mode/2up

an analytic pipeline made it possible to trace each modeling step with precision. Our aim is not to reassess the study's substantive conclusions but to clarify how specific analytic choices influence the fitted surface and how guidance from the methodological literature can help researchers navigate these decisions. In doing so, we hope to contribute to broader conversations about methodological transparency and to support more robust practices in the applied use of complex analytical techniques in educational research.

2. Case Study Context: A Response Surface Analysis of Stereotype Threat Effects in Education

Leroy et al. [4] set out to examine the extent to which a commonly held gender stereotype in the school context—specifically, the belief that boys are better at math than girls—affects students' academic self-perceptions and outcomes. Drawing on prior work, they distinguished between "stereotype knowledge" (being aware of the stereotype) and "stereotype belief" (endorsing its validity) [11]. Earlier studies had shown that both forms of endorsement could negatively impact students, especially girls, and that their combination might further intensify these effects [12, 13].

While the theoretical basis of the study was built on well-established ideas, its methodological novelty lay in the use of RSA to explore how stereotype knowledge and belief might interact. RSA enables researchers to examine how an outcome varies across combinations of two predictor variables, focusing in particular on the line of congruence (LOC), where predictors are equal, and the line of incongruence (LOIC), where they differ most. In this case, RSA was used to model how math self-efficacy and achievement varied depending on students' levels of stereotype knowledge and belief.

The technique relies on polynomial regression to allow both linear and nonlinear patterns in this joint influence. The quadratic polynomial model $z = b_0 + b_1X + b_2Y + b_3X^2 + b_4XY + b_5Y^2$ is commonly used for this purpose because its coefficients can be rearranged to describe the outcome's behavior along the LOC and the LOIC, notably with the function $z(LOC) = b_0 + a_1X + a_2X^2$ and the function $z(LOIC) = b_0 + a_3X + a_4X^2$, where $a_1 = b_1 + b_2$, $a_2 = b_3 + b_4 + b_5$, $a_3 = b_1 - b_2$, $a_2 = b_3 - b_4 + b_5$ [5]. These composite parameters, rather than the individual polynomial terms, summarize the slope and curvature of the surface in the regions that correspond to the researcher's hypotheses and therefore provide the main basis for interpretation. RSA also involves projecting the fitted surface to examine its geometry in relation to the expected pattern, for example, whether the outcome increases when the predictors align or whether curvature suggests diminishing or amplifying effects at extreme combinations.

The dissemination work by Humberg et al. [5] has made RSA more accessible to applied researchers by providing detailed introductions, ready-to-use scripts, and a dedicated R package with graphical capabilities. These resources have encouraged broader use of RSA and made its application more intuitive. For example, in their description of the analytic procedures, Leroy et al. [4] noted that their RSA was based on scripts developed by Humberg et al. [5], which offer detailed guidance and visualization tools that have greatly enhanced the accessibility of RSA in applied research. At the same time, the apparent simplicity of these tools can obscure analytic decisions that are central to the method. Choices concerning centering or standardization, treatment of missing data, comparison of polynomial levels, and interpretation of surface parameters are not technical details but integral parts of the modeling framework. As the sections below describe, several aspects of the implementation in the case study illustrate how these steps influence the shape and interpretation of the fitted surface and why they require careful consideration in applied RSA.

3. Analytic Approach

To examine the methodological issues arising in the application of RSA in the case study, we conducted a close review of the research and of the publicly available analysis scripts and data. We systematically reconstructed the modeling pipeline, comparing stated analytic intentions with actual implementation choices in the code and reported outputs. Our evaluation drew on established RSA guidelines [5, 6, 14], with a focus on specification alignment, data handling, variable transformation, and model interpretation. When discrepancies were detected between reported procedures and the underlying statistical logic, we cross-validated these against recommendations from the methodological literature. Where appropriate, we referenced alternative modeling strategies that could have clarified or corrected the analytic claims. This structured approach was intended not only to identify common pitfalls in an illustrative study case but also to distill broader lessons for applied research using RSA (Table 1).

Table 1
Summary of methodological issues and possible corrections

Methodological issue	What the case study did	Why this is an issue	Correction
1. Mismatch between research hypotheses and RSA specification	Hypotheses were framed as linear, yet quadratic RSA was applied (Table 2)	Quadratic RSA addresses hypotheses about nonlinear (quadratic) processes	Introduce hypotheses positing nonlinear effects of predictors on the outcomes
2. Mismatch between reported missing data treatment and actual specification appearing in the RSA scripts	Analyses were conducted with listwise deletion, despite the methods section stating FIML was used	Listwise deletion and FIML are two incompatible and contradictory ways of treating missing data	Replace the statement "na.rm = TRUE" by "na.rm = FALSE, missing = 'fiml' " to account for missing data using the FIML estimation procedure

(*Continued*)

Table 1
(*Continued*)

Methodological issue	What the case study did	Why this is an issue	Correction
3. Confusion of z-standardization and grand-mean centering in RSA	Predictors were described as both z-standardized and grand-mean centered	Standardization to z scores and grand-mean centering are two opposite ways of transforming variables	Choose and justify one and only standardization procedure for the analysis (e.g., z-standardization, grand-mean centering, grand standardization)
4. Inappropriate sensitivity analysis	Sensitivity analysis was reported for the full sample, though RSA was conducted separately for boys and girls	The sensitivity analysis should be conducted on the same sample as the one used in the RSA	Conduct sensitivity analyses separately for all samples (or the smallest sample) used in the final RSA
5. Lack of testing for differences between boys and girls	RSA differences between boys and girls were interpreted as reflecting gender variations in the processes, without tests	Group differences can only be inferred using formal tests for measurement and structural differences	Compare single-group and multi-group (boys vs. girls) solutions using likelihood ratio test (LRT) and fit indices
6. Overreliance on single RSA parameters to interpret the response surfaces	Interpretation relied on individual regression terms rather than composite surface parameters, leading to interpretive error (Figure 1)	Response surfaces are a byproduct of all polynomial effects, isolated inferences may be misleading	Characterize the response surface accurately by identifying meaningful curvatures in the surface (e.g., extrema)
7. Lack of assessment of model fit required for RSA	Final models for RSA were based only on the fact that they significantly differed from an empty model (omnibus F test)	Nonlinear effects should be tested before concluding the existence of nonlinear processes	Test the relevance nonlinear effects by comparing nested or non-nested models using comparative fit indices

Note: Explanations and corrections are based on applied statistics and RSA methodology.

4. Dissecting the Knowledge Gap: Methodological Issues in RSA and Ways to Rectify Them

We now examine a series of methodological issues—conceptual, operational, interpretive, and statistical—that emerged in the application of RSA by Leroy et al. [4]. While based on a single study, the issues reflect broader challenges that may arise, particularly when advanced statistical techniques are implemented using pre-written scripts and key analytical decisions are guided more by technical convenience than by close alignment with the underlying research questions.

Table 1 summarizes the main concerns identified, their implications, and proposed corrections. These include model misspecification, unclear data handling procedures, and interpretive inconsistencies. Each issue is discussed in the sections that follow, with the aim of clarifying how similar difficulties might be recognized and addressed in other contexts.

4.1. Mismatch between research hypotheses and RSA specification (issue 1)

Although RSA can be used to examine both linear and nonlinear relations, the interpretation of a quadratic model depends on whether second-order patterns are anticipated theoretically. Linear expectations concern directional effects along the line of congruence or incongruence, whereas quadratic models imply expectations about curvature in these regions [5, 6]. In the case study, however, all four hypotheses expressed linear expectations—such as higher self-efficacy when belief is lower than knowledge or when both are low—without specifying any nonlinear components. As shown in Table 2, these expectations emphasize first-order relations and therefore do not align with the quadratic model that was estimated.

The only partial exception is Hypothesis 3, which refers to "the degree of discrepancy in favor of belief." This could be taken to suggest a curvilinear effect along the LOIC, yet the wording remains ambiguous. As phrased, H3 appears to restate H2 in different terms rather than positing a distinct, second-order effect. If the intention was to test whether outcomes become increasingly negative at higher levels of discrepancy, this would require a clearer, explicitly quadratic hypothesis.

This mismatch makes it difficult to align the estimated surface with the stated hypotheses and introduces a risk of misinterpretation of the results. To avoid this, future studies should ensure that the statistical model reflects the theoretical claims being tested. When using quadratic RSA, this means formulating hypotheses that include possible nonlinearities, such as threshold effects at extreme values of congruence or discrepancy [6, 15].

4.2. Mismatch between reported missing data treatment and actual specification appearing in the analysis scripts (issue 2)

In the "Data preparation" section, the case study stated that missing data were handled using full information maximum likelihood (FIML), noting that this approach reduces bias, particularly when missingness affects squared or interaction terms. This concern is well established, as polynomial models depend on the joint behavior of X, Y, X^2, XY, and Y^2, which are sensitive to distortions in the covariance structure. FIML operates by estimating model parameters from all available data points under the assumption that data are missing at random and does so

Table 2
Mismatch between study hypotheses and use of quadratic RSA

Hypothesis	Model specification	Response surface
H1: "We expected math self-efficacy and performance to be higher for girls when knowledge and belief of the negative stereotype aligned at a low level [...] rather than at a high level."	$z = b_0 + b_1X + b_2Y + e$ with: $a_1 = b_1 + b_2 > 0$	Linear (not quadratic)
H2: "We expected math self-efficacy and performance to be higher for girls when stereotype belief is lower than stereotype knowledge [...] rather than when stereotype belief is higher than stereotype knowledge."	$z = b_0 + b_1X + b_2Y + e$ with: $a_3 = b_1 - b_2 < 0$	Linear (not quadratic)
H3: "We expected girls' math self-efficacy and performance to decrease as the degree of the discrepancy in favor of belief increased."	Same as H2	Linear (not quadratic)
H4: "We expected that when stereotype knowledge and stereotype belief were aligned at a high level, girls' self-efficacy and math performance would be lower than when stereotype belief was lower than stereotype knowledge [...]. Additionally, we expected math self-efficacy and performance to be higher when stereotype belief is higher than stereotype knowledge [...] compared to when stereotype knowledge and stereotype belief are aligned at a high level."	$z = b_0 + b_1X + b_2Y + e$ with: $a_1 = b_1 + b_2 > 0$ $a_3 = b_1 - b_2 < 0$ $\lvert a_3 \rvert > \lvert a_1 \rvert$	Linear (not quadratic)

Note: The general form of the equation to test for quadratic relationships in RSA using polynomial regression is $Z = b_0 + b_1X + b_2Y + b_3X^2 + b_4XY + b_5Y^2 + e$, where Z is a dependent variable, X is predictor 1 (stereotype knowledge), and Y is predictor 2 (stereotype belief). None of the hypotheses imply second-order effects (i.e., b_3, b_4, or b_5) that are characteristic of quadratic RSA.

without discarding cases [16, 17]. However, the R scripts associated with the study specified na.rm = TRUE, which performs listwise deletion. Because listwise deletion removes any case with a missing value on any variable, it modifies the covariance matrix on which polynomial models depend and can lead to parameter estimates that differ systematically from those obtained with FIML. These two approaches—listwise deletion and FIML—rely on incompatible assumptions and therefore cannot be used interchangeably.

This discrepancy affects both the reproducibility and interpretability of the resulting response surface (see 4.8 for an illustration). Unfortunately, it is not possible to assess how the use of listwise deletion influenced the results of the case study, because the data currently available on the study's OSF repository have already been reduced to complete cases—198 girls and 255 boys—whereas the reported sample consisted of 294 girls and 272 boys. In other words, the OSF data reflect the outcome of listwise deletion and cannot be used to re-estimate the models using FIML.

This pitfall could have been avoided by adapting the R scripts (retrieved from Humberg et al. [5]) and specifying na.rm = FALSE to retain the missing data for analysis, in combination with estimation procedures compatible with FIML[2].

4.3. Confusion of z-standardization and grand-mean centering (issue 3)

In their description of data preparation, Leroy et al. [4] stated that both predictors were "z-standardized based on the grand mean" and subsequently noted that they centered the variables at the grand mean. This phrasing conflates two distinct transformations. Z-standardization rescales each variable individually to have a mean of zero and a standard deviation of one, whereas grand-mean centering subtracts a single shared mean from both variables without altering their variances. The former is a within-variable transformation; the latter is a between-variable alignment procedure intended to place predictors on a common reference point. Treating these procedures as interchangeable makes it unclear how the variables were prepared and complicates the interpretation of the polynomial components.

These distinctions are not merely terminological. RSA relies on squared and interaction terms whose magnitude depends directly on the scaling of the predictors. As a result, different transformations can produce surfaces that differ in shape, curvature, and orientation. When it is unclear how the predictors were transformed, the interpretive meaning of the resulting surface parameters becomes uncertain, particularly for the curvature components derived from the quadratic terms.

A brief numerical illustration may help clarify this point. Suppose three hypothetical observations for stereotype knowledge and stereotype belief are (4, 6), (5, 5), and (6, 4). Grand-mean centering subtracts the same overall mean ($M = 5$) from both variables, producing centered values of (−1, 1), (0, 0), and (1, −1). Z-standardization, by contrast, rescales each variable using its own mean and standard deviation, yielding values that

[2]The change to the OSF script—replacing na.rm = TRUE with na.rm = FALSE—was introduced on March 16, 2025 (according to the blinded OSF repository), shortly after the authors were granted access to an earlier version of this manuscript during the peer-review process. However, because the repository was shared via a view-only link (https://osf.io/s57nb/?view_only=be45e71d9442410caf1e9bdd5b5d4170), no version history is preserved, and the original script containing na.rm = TRUE is no longer publicly accessible. The dataset uploaded alongside this revision had already been listwise-deleted, rendering the script modification inoperant and potentially misleading. To safeguard against further loss of transparency, we provide the original script and data used by the authors via the following archival OSF repository: [https://osf.io/fcqr2/overview?view_only=82fe0a78a36e4993b03a676c5e14638b].

differ both in location and spread. The squared and interaction terms derived from these two sets of transformed predictors are therefore not equivalent. Because the curvature of the response surface depends on these higher-order terms, different transformations can lead to different surface shapes (see 4.8 for an illustration). Clear reporting of the chosen procedure is therefore essential.

As the shared analysis script contained data that had already been listwise-deleted and transformed, it is not possible to evaluate whether the choice of standardization influenced the reported findings. This pitfall could have been avoided by applying a clearly defined transformation aligned with RSA recommendations (e.g., z-standardization or grand standardization) and preserving raw input data to allow others to evaluate its impact on the analysis.

4.4. Inappropriate sensitivity analysis (issue 4)

In the "Participants" section, the case study reported conducting a sensitivity analysis indicating that, with 566 participants, their design could detect effects as small as $d = 0.16$. However, the RSA models were fitted separately for girls and boys, using subsamples of 294 and 272 participants, respectively. This introduces a mismatch between the reported sensitivity and the actual sample sizes used in the analyses. As a result, the analyses had approximately half the statistical power reported by the authors.

This issue is particularly relevant in RSA because polynomial models consume more degrees of freedom than linear models and therefore require larger samples to detect curvature or asymmetry in the response surface. When the sample size is overestimated, the risk increases that true nonlinearities remain undetected, producing null findings that may reflect limited power rather than the absence of nonlinear associations [5]. Indeed, several of the models reported by Leroy et al. [4] did not yield significant curvature effects, especially for achievement outcomes, making it difficult to determine whether these null results reflect genuine linearity or insufficient power.

This pitfall could have been avoided by aligning the sensitivity analysis with the analytic subsamples or by reporting power estimates separately for each group, which would have provided a more accurate basis for interpreting the results.

4.5. Lack of testing for differences in RSA between boys and girls (issue 5)

Leroy et al. [4] estimated separate RSA models for girls and boys and interpreted apparent differences in the shape and strength of the surfaces. However, no formal test was reported to evaluate whether the model parameters differed significantly across groups. This is concerning because apparent differences between groups, be it in terms of direction or significance of effects, may simply reflect non-informative, random sampling variance or measurement error [18]. Without explicit tests, it is not possible to determine whether the observed differences reflect true group-level variation or are simply due to random variation in parameter estimates.

In the context of RSA, assessing group differences requires evaluating whether the polynomial coefficients—and therefore the surface parameters—differ across groups in a statistically meaningful way. This typically involves estimating multigroup models (boys vs girls) with and without invariant path coefficient constraints and comparing their relative fit, as illustrated in previous applications [19]. Formal tests of this kind would clarify whether the reported differences in responses to stereotype knowledge and belief reflect genuine gender-related variation or fall within the range of sampling variability.

4.6. Overreliance on single parameters to interpret response surfaces (issue 6)

In reporting their results, the case study emphasized the significance or direction of individual regression coefficients when interpreting the response surfaces. For instance, they highlighted the presence or absence of isolated polynomial terms (e.g., b_3, b_4) to describe whether the surface exhibited curvature or congruence effects. However, interpretations based on isolated terms can misrepresent the surface as a whole.

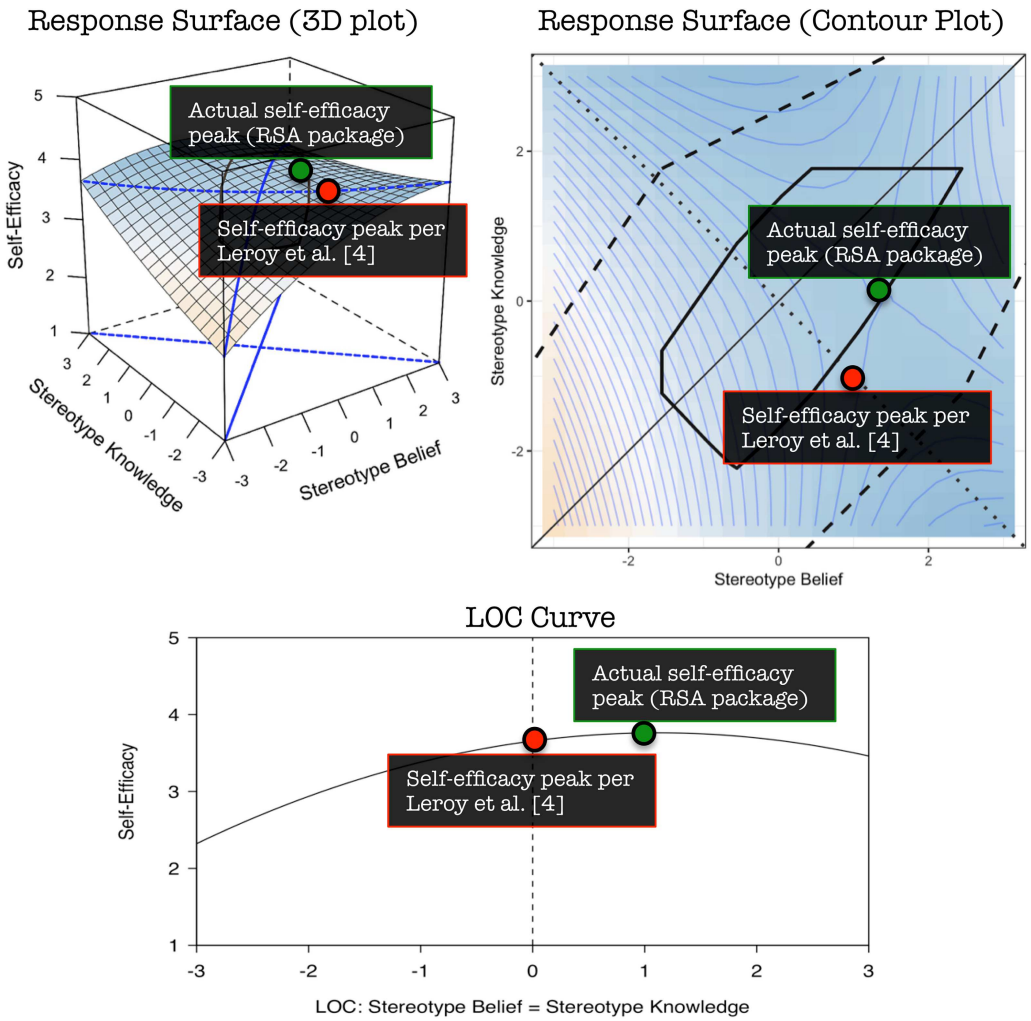
RSA methodology has consistently emphasized the importance of considering all parameters in a polynomial regression model before interpreting the response surface, notably by identifying extrema that reflect the joint behavior of the predictors [5, 15]. This is because the surface arises from the combined influence of all coefficients shaping its curvature (e.g., b_3, b_4, or b_5) and location (i.e., b_1, b_2). Only when certain parameters are small can they be disregarded—and even then, this requires formal testing [20].

By contrast, the study by Leroy et al. [4] interpreted response surfaces based on isolated polynomial terms. For example, in the model predicting girls' self-efficacy, the authors noted that the curvature along the LOC was negative ($a_2 = -0.06$) and concluded that alignment between belief and knowledge had a detrimental nonlinear effect on self-efficacy. This interpretation was misleading because the quadratic effect of the LOC $a_2 = b_3 + b_4 + b_5$ actually reduced to the quadratic effect of stereotype belief $b_3 = -0.07$, which was the only significant and nontrivial effect (i.e., b_4 and b_5 were close to null). In other words, the apparent nonlinear effect of alignment was illusory, as it only reflected variations in stereotype belief (not in alignment).

A similar issue arises in the model for boys' self-efficacy (see Figure 1). There, the study interpreted the negative interaction term ($b_4 = -0.05$) as indicating that high knowledge dampens the positive effect of belief. This led them to infer that the highest self-efficacy levels would be found when belief is high but knowledge is low or, stated otherwise, with predictor values situated near 1 SD above the mean for belief and 1 SD below the mean for knowledge. However, when using all parameters to locate the extremum of the surface, the vertex is found at approximately $(X, Y) = (1.3, 0.1)$—that is, above-average levels of both belief and knowledge. Likewise, the negative quadratic effect on the LOC (i.e., $a_2 = -0.085$) was used to suggest that alignment at high levels leads to lower self-efficacy. However, this interpretation overlooks that the positive slope along the LOC (i.e., a_1=0.19) predicts an increasing trend, shifting the extremum to $X_{ext} = \frac{a_1}{-2a_2} = \frac{0.19}{-2*-0.085} = 1.12$ [15]. The surface predicts maximum self-efficacy not at low levels of belief and knowledge—as the authors implied—but at high, aligned values on both predictors, contradicting their interpretation (Figure 1).

These misinterpretations resulted from considering parameters in isolation. They could have been avoided by examining the response surface as a whole, using composite surface parameters (a_1 to a_4) that summarize slopes and curvatures along the lines of congruence and incongruence. These parameters, together with visual inspection of the surface and calculation of extrema, offer a more reliable basis for interpretation in RSA.

Figure 1
Misinterpretations due to overreliance on single parameters



Note: Figure reproduced from the case study data. Peaks in red indicate the highest levels of self-efficacy misinterpreted due to overreliance on single polynomial parameters, while peaks in green show the true highest levels computed by incorporating all parameters, as provided by the R package *RSA*.

4.7. Lack of assessment of model fit required for RSA (issue 7)

Final models for RSA in the case study were validated only based on the observation that they significantly differed from an empty model (omnibus F test). However, RSA relies on polynomial components whose contribution must be evaluated relative to simpler specifications. In this context, omnibus tests do not indicate whether the inclusion of quadratic terms improves the representation of the data or whether the observed surface features reflect meaningful structure rather than sampling variation. Nonlinear effects should be tested before concluding the existence of nonlinear processes, especially because curvature along the LOC and LOIC depends directly on the second-order parameters.

In the present case, several of the reported polynomial effects were nonsignificant [4], which raises questions about whether the estimated nonlinearities were necessary or whether a simpler linear model would have sufficed. Best practices in RSA recommend assessing whether a more complex polynomial model (e.g., including quadratic or cubic terms) provides a significantly better fit than a simpler model (e.g., linear model), for example, by examining ΔR^2 or chi-square difference tests for nested models, and information criteria such as Akaike information criterion (AIC) or Bayesian information criterion (BIC) for broader comparisons [5–7, 14, 15]. Without such comparisons, it is difficult to judge whether the selected model is appropriately specified or whether its curvature is an artifact of overfitting or noise.

4.8. Simulation example of how analytic decisions can distort RSA results

To illustrate how analytic decisions can influence the behavior of an RSA, we conducted a small simulation based on the response surface for girls' self-efficacy, which was the clearest pattern reported in the case study[3]. Using the polynomial coefficients reported in the work from Leroy et al. [4] as population-generating values, we first simulated stereotype belief and stereotype knowledge as correlated predictors (r = 0.43), matching the distribution reported for girls. Self-efficacy scores were then generated

[3]The simulation results can be reproduced here: https://osf.io/fcqr2/overview?view_only=82fe0a78a36e4993b03a676c5e14638b.

from the quadratic model using these predictors, with random error added to approximate the variability observed in the empirical data. This dataset represents the "true" process, in which all cases are complete, predictors retain their original scales, and the RSA is estimated using all available information (Figure 2, Panel A). From this baseline, we examined how the fitted surface changes when applying three analytic modifications that mirror issues discussed earlier.

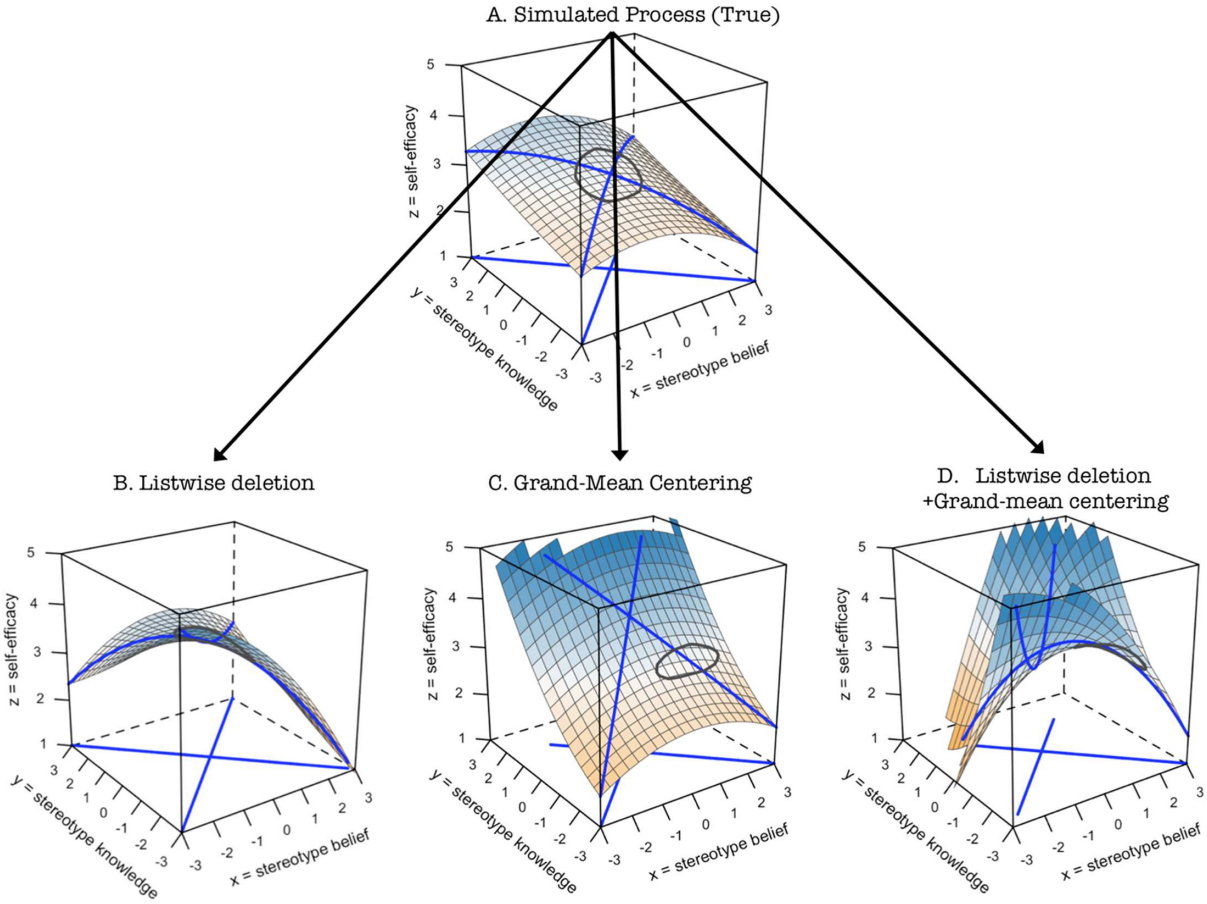
First, to reflect the consequences of listwise deletion, we imposed missing data using a simple missing-at-random mechanism in which observations with values above approximately +1 SD on either predictor had a higher probability of being removed. We then estimated the RSA using only the remaining complete cases. As shown in Figure 2 (Panel B), the resulting surface deviated substantially from the simulated process, with noticeable shifts in curvature.

Second, to illustrate the effect of grand-mean centering, we analyzed the complete simulated dataset after centering the raw predictors on their pooled mean before constructing the polynomial terms. This transformation altered the scale and alignment of the predictors, leading to changes in the curvature and orientation of the response surface (Figure 2, Panel C). Because squared and interaction terms in RSA are sensitive to the scale of the underlying predictors, such changes directly affect the behavior of the fitted surface.

Finally, we combined these two modifications by imposing missingness as above and then applying grand-mean centering (Figure 2, Panel D). This produced the largest deviations from the "true" surface, with both curvature and the predicted maximum shifting markedly. In this scenario, the overall shape of the original surface was no longer readily recognizable.

Taken together, this simulation shows that the handling of missing data and the choice of standardization or centering procedures can meaningfully distort the geometry of an RSA surface. These findings do not imply that the results of the original study are necessarily affected in the same way. Because the publicly available dataset in the work of Leroy et al. [4] has already undergone listwise deletion and transformation, it is not possible to assess how these analytic choices influenced their reported surfaces. The simulation serves only to illustrate how such decisions can shape RSA results and why transparent reporting of each step is important in applied research.

Figure 2
Response surfaces obtained under different analytic conditions in the simulation study



Note: Panels illustrate the response surface implied by the simulated "true" process (A), together with surfaces estimated after listwise deletion (B), grand-mean centering (C), and the combination of both procedures (D). The simulation uses the polynomial coefficients for girls' self-efficacy reported in the case study [4] as the population-generating model. The comparison shows how missing data handling and centering decisions can alter the shape, curvature, and extremum of the fitted surface. Because the publicly available data from the original study had already undergone listwise deletion and transformation, it is not possible to determine whether similar distortions occurred in their analyses; the figure is intended only to illustrate how such analytic choices may have influenced RSA results.

5. Discussion

The present case reflection has examined a series of methodological issues that arose in the application of RSA in a recent educational study. These included inconsistencies between analytic goals and implementation, unclear or incorrect treatment of missing data and variable transformations, and interpretations that relied too narrowly on individual model terms. While each point may seem modest on its own, their combined effect raises concerns about how reliably the results can be interpreted—especially in methods like RSA, where different parameters interact in complex ways [5, 6, 15]. For instance, our simulation example based on the case study (Figure 2) illustrated that routine analytic decisions such as listwise deletion or certain centering transformations can meaningfully alter the geometry of a fitted surface, suggesting that even modest deviations from recommended procedures may have practical consequences for interpretation.

These concerns do not mean that the study's conclusions are necessarily wrong, but that the evidence is itself inconclusive. They also reflect broader challenges faced by educational researchers when using advanced statistical techniques—especially when those techniques are applied through open-access scripts. Even when such scripts are well-documented, they may be used in ways that diverge from their intended purpose if the underlying assumptions or decision points are not fully understood. This can result in analyses that appear technically correct but rest on unclear or inconsistent modeling choices [5].

While the present reflection centers on a single case, similar methodological limitations can be observed in other recent educational studies applying RSA. Other studies by Leroy et al. [4, 21, 22] illustrate similar modeling and interpreting challenges. Some of these challenges can also be observed in independent studies unrelated to these authors. For example, Bai et al. [8], Lan et al. [9], and Lee et al. [10] each report RSA findings in which the response surface is interpreted without fully characterizing its curvature (issue 6) or without reporting model comparisons or tests of whether nonlinear terms significantly improve model fit (issue 7). These examples suggest that the challenges illustrated here are not isolated, but part of a broader pattern in the applied use of RSA—especially when implementation relies on shared scripts rather than a full understanding of the modeling framework.

In this respect, the issues discussed here speak to a broader, and still emerging, conversation about open science practices in education [23, 24]. As researchers increasingly adopt open tools and share analysis scripts, there is growing recognition that transparency alone is not enough. In the case examined here, scripts and materials were indeed made available—but the way they were applied left key analytic decisions—such as how data were prepared, models specified, or results interpreted—unclear or inconsistent. As open science gains traction across fields [1, 2], the ability to share code and data is a valuable step forward. Yet it must be matched by careful modeling choices and clear alignment between theory, method, and interpretation.

We hope this reflection contributes to that shift by illustrating how openness can be strengthened through methodological clarity and thoughtful implementation. In particular, the structured summary provided in Table 1 may serve as a practical reference point for researchers and reviewers seeking to avoid similar pitfalls. In this perspective, we see this reflection as contributing to a critical examination of how analytic choices shape, and sometimes obscure, the knowledge claims made in educational research. By drawing attention to these often-implicit modeling assumptions and by demonstrating their potential consequences in a simple simulated setting, we hope to support a more reflective use of quantitative methods—one that remains mindful of their technical demands.

6. Conclusion

Application of RSA, even when supported by open-access scripts, can present challenges when analytic decisions are not closely aligned with the underlying modeling assumptions. We have sought to clarify how issues of model specification, data treatment, and interpretation may interact in ways that merit closer scrutiny. While the discussion is grounded in a single case, the analytic patterns it highlights are not unique, and they point to broader questions about the implementation of complex statistical tools in educational research [25, 26]. We hope this contribution supports ongoing conversations about methodological transparency and helps foster greater attentiveness to the interpretive demands of advanced modeling techniques as they become more widely used in applied settings.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in OSF at https://osf.io/fcqr2/overview?view_only=82fe0a78a36e4993b03a676c5e14638b.

Author Contribution Statement

Kim Sorjonen: Conceptualization, Methodology, Validation, Formal Analysis, Writing – original draft, Writing – review and editing; **Rebeka Weidmann:** Conceptualization, Methodology, Validation, Writing – review and editing.

References

[1] Miyakawa, T. (2020). No raw data, no science: Another possible source of the reproducibility crisis. *Molecular Brain*, *13*(1), 24. https://doi.org/10.1186/s13041-020-0552-2

[2] Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., . . . , & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

[3] R Core Team. (2025). *R: A language and environment for statistical computing (Version 4.5) [Computer software]. The R Foundation*. Retrieved from: https://www.R-project.org

[4] Leroy, N., Max, S., & Pansu, P. (2025). How stereotype knowledge and stereotype belief impact girls' self-efficacy and math performance: A response surface analysis approach. *Learning and Instruction*, *96*, 102071. https://doi.org/10.1016/j.learninstruc.2024.102071

[5] Humberg, S., Nestler, S., & Back, M. D. (2019). Response surface analysis in personality and social psychology: Checklist and clarifications for the case of congruence hypotheses. *Social Psychological and Personality Science*, *10*(3), 409–419. https://doi.org/10.1177/1948550618757600

[6] Humberg, S., Schönbrodt, F. D., Back, M. D., & Nestler, S. (2022). Cubic response surface analysis: Investigating asymmetric and level-dependent congruence effects with third-order polynomial models. *Psychological Methods*, *27*(4), 622–649. https://doi.org/10.1037/met0000352

[7] Shanock, L. R., Baran, B. E., Gentry, W. A., Pattison, S. C., & Heggestad, E. D. (2010). Polynomial regression with response surface analysis: A powerful approach for examining moderation and overcoming limitations of difference scores. *Journal of Business and Psychology*, *25*(4), 543–554. https://doi.org/10.1007/s10869-010-9183-4

[8] Bai, X., Song, Y., Liang, M., Peng, J., Ruan, T., & Xie, H. (2025). Relationship between parent-adolescent discrepancies in educational expectations and academic burnout: A response surface analysis. *Acta Psychologica*, *253*, 104766. https://doi.org/10.1016/j.actpsy.2025.104766

[9] Lan, Z., Liu, H., Huang, X., Wang, Q., Deng, F., & Li, J. (2023). The impact of academic pressure and peer support on adolescents' loneliness: A polynomial regression and response surface analysis. *Psychology Research and Behavior Management*, *16*, 4617–4627. https://doi.org/10.2147/PRBM.S435977

[10] Lee, B., Park, H. I., & Park, S. (2022). Parent-adolescent vocational aspiration congruence and its relations with academic adjustment. *Journal of Youth and Adolescence*, *51(*7) 1374–1387. https://doi.org/10.1007/s10964-022-01598-1

[11] Régner, I., Steele, J. R., Ambady, N., Thinus-Blanc, C., & Huguet, P. (2014). Our future scientists: A review of stereotype threat in girls from early elementary school to middle school. *Revue Internationale de Psychologie Sociale*, *27*(3–4), 13–51.

[12] Bonnot, V., & Croizet, J.-C. (2011). Stereotype threat and stereotype endorsement: Their joint influence on women's math performance. *Revue Internationale de Psychologie Sociale*, *24*(2), 105–120.

[13] Schmader, T., Johns, M., & Barquissau, M. (2004). The costs of accepting gender differences: The role of stereotype endorsement in women's experience in the math domain. *Sex Roles*, *50*(11), 835–850. https://doi.org/10.1023/B:SERS.0000029101.74557.a0

[14] Edwards, J. R. (2002). Alternatives to difference scores: Polynomial regression analysis and response surface methodology. In F. Drasgow & N. Schmitt (Eds.), *Measuring analyzing behavior organizations: Advances in measurement and data analysis* (pp. 350–400). Jossey-Bass/Wiely.

[15] Edwards, J. R., & Parry, M. E. (1993). On the use of polynomial regression equations as an alternative to difference scores in organizational research. *Academy of Management Journal*, *36*(6), 1577–1613. https://doi.org/10.2307/256822

[16] Enders, C. K. (2025). Missing data: An update on the state of the art. *Psychological Methods*, *30*(2), 322–339. https://doi.org/10.1037/met0000563

[17] Rioux, C., & Little, T. D. (2021). Missing data treatments in intervention studies: What was, what is, and what should be. *International Journal of Behavioral Development*, *45*(1), 51–58. https://doi.org/10.1177/0165025419880609

[18] Leitgöb, H., Seddig, D., Asparouhov, T., Behr, D., Davidov, E., de Roover, K., . . . , & van de Schoot, R. (2023). Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. *Social Science Research*, *110*, 102805. https://doi.org/10.1016/j.ssresearch.2022.102805

[19] Weidmann, R., Schönbrodt, F. D., Ledermann, T., & Grob, A. (2017). Concurrent and longitudinal dyadic polynomial regression analyses of Big Five traits and relationship satisfaction: Does similarity matter? *Journal of Research in Personality*, *70*, 6–15. https://doi.org/10.1016/j.jrp.2017.04.003

[20] Weidmann, R., Alabdullah, A., Warkentien, J., Vesikar, S., Denehy, M. E., Yang, H., . . . , & Lucas, R. E. (2024). Does perceived similarity between present and future self predict well-being? A reanalysis and replication of Reiff et al. *Social Psychological and Personality Science*, *15*(7), 791–801. https://doi.org/10.1177/19485506241257971 2020

[21] Leroy, N. (2025). Professional development, teacher job satisfaction, and work engagement: A needs-supplies fit approach. *Professional Development in Education*, 1–22. https://doi.org/10.1080/19415257.2025.2550987

[22] Leroy, N. (2025). Testing the effect of agreement and discrepancy between teachers' and students' reported interpersonal behaviors on students' motivation: A response surface analysis. *Social Psychology of Education*, *28*(1), 136. https://doi.org/10.1007/s11218-025-10096-8

[23] Abid, M. N., Malik, A., & Sarwar, S. (2023). Research reproducibility ethics of scientific research in higher education. *Journal of Educational Research and Social Sciences Review*, *3*(3), 37–50.

[24] Gehlbach, H., & Robinson, C. D. (2021). From old school to open science: The implications of new research norms for educational psychology and beyond. *Educational Psychologist*, *56*(2), 79–89. https://doi.org/10.1080/00461520.2021.1898961

[25] Sorjonen, K., Melin, B., & Melin, M. (2025). Questionable necessity effects of basic psychological needs at work: A comment on Ding and Kuvaas. 2025 *PsyArXiv*. https://doi.org/10.31234/osf.io/wneu2_v1

[26] Sorjonen, K., & Melin, B. (2025). Adding some rigor to necessary condition analysis. *PsyArXiv*. https://doi.org/10.31234/osf.io/sq3b9_v1

How to Cite: Sorjonen, K., & Weidmann, R. (2026). Exploring Methodological Challenges in Educational Research: A Case Reflection on Response Surface Analysis. *International Journal of Changes in Education*. https://doi.org/10.47852/bonviewIJCE62027911