



## UvA-DARE (Digital Academic Repository)

### **Absence of evidence is not evidence of absence**

*On the limited use of regression discontinuity analysis in higher education*

van Dorresteijn, C.; Kan, K.-J.; Smits, N.

#### **DOI**

[10.1080/02602938.2021.2016606](https://doi.org/10.1080/02602938.2021.2016606)

#### **Publication date**

2023

#### **Document Version**

Final published version

#### **Published in**

Assessment and Evaluation in Higher Education

#### **License**

CC BY-NC-ND

[Link to publication](#)

#### **Citation for published version (APA):**

van Dorresteijn, C., Kan, K.-J., & Smits, N. (2023). Absence of evidence is not evidence of absence: On the limited use of regression discontinuity analysis in higher education. *Assessment and Evaluation in Higher Education*, 48(1), 16-26. Advance online publication. <https://doi.org/10.1080/02602938.2021.2016606>

#### **General rights**

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

#### **Disclaimer/Complaints regulations**

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

# Absence of evidence is not evidence of absence. On the limited use of regression discontinuity analysis in higher education

Chevy van Dorresteijn, Kees-Jan Kan & Niels Smits

**To cite this article:** Chevy van Dorresteijn, Kees-Jan Kan & Niels Smits (2023) Absence of evidence is not evidence of absence. On the limited use of regression discontinuity analysis in higher education, *Assessment & Evaluation in Higher Education*, 48:1, 16-26, DOI: [10.1080/02602938.2021.2016606](https://doi.org/10.1080/02602938.2021.2016606)

**To link to this article:** <https://doi.org/10.1080/02602938.2021.2016606>



© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 21 Dec 2021.



Submit your article to this journal [↗](#)



Article views: 1309






View related articles [↗](#)



View Crossmark data [↗](#)

# Absence of evidence is not evidence of absence. On the limited use of regression discontinuity analysis in higher education

Chevy van Dorresteijn , Kees-Jan Kan  and Niels Smits 

Research Institute of Child Development and Education, University of Amsterdam, Amsterdam, The Netherlands

## ABSTRACT

When higher education students are assessed multiple times, teachers need to consider how these assessments can be combined into a single pass or fail decision. A common question that arises is whether students should be allowed to take a resit. Previous research has found little to no clear learning benefits of resits and therefore suggested they might not be advantageous as they are costly for both students and institutions. However, we conducted a simulation study that shows such a conclusion to be presumptuous. Absence of evidence is not evidence of absence; our results illustrate that if a resit effect were to exist, the analysis used in these studies (i.e. regression discontinuity analysis; RDA) lacked the power to detect such an effect. Power of RDA was only sufficient under extremely implausible conditions (i.e. large sample, large effect size, high correlation between examinations). To adequately compare the effect of assessment policies, researchers are recommended to use other methods than RDA.

## KEYWORDS

regression discontinuity analysis; simulation study; statistical power; resit effect

The COVID-19 pandemic forced a rapid, unprecedented transition from on-campus to distant learning that required educational institutions to reconsider their assessment practices (Prigoff, Hunter, and Nowygrod 2021; Tan et al. 2021). Examples of modified assessment procedures include the replacement of closed-ended items by open-ended questions, on-campus examinations by take-home assignments, and a single final examination by multiple assessments. Importantly, while such ad hoc changes are completely understandable during a pandemic, they should not affect the educational standards. This might be quite a challenge, since different assessment methods have different aims (Walstad 2001). For instance, closed-ended items are used to assess students' levels of knowledge, while take-home assignments may be better suited to assess reflective thinking. Further, although a single examination and multiple assessments may target the same knowledge domain, multiple assessments have been typically recommended to monitor students' learning process (Biggs and Tang 2007).

If multiple assessments are used, a decision has to be made on how to combine them into a single pass or fail decision. Commonly, this comes down to choosing between two types of decision rules: a conjunctive rule or a compensatory rule. A conjunctive rule requires students to obtain a satisfactory grade for each assessment to pass a course. A compensatory rule, by contrast, allows students to use satisfactory grades to compensate for unsatisfactory grades as long as the combined grade meets the standard. The debate as to whether one rule is better than the other is relevant beyond the COVID-19 pandemic and has a long history. The

**CONTACT** Chevy van Dorresteijn  [c.m.vandorresteijn@uva.nl](mailto:c.m.vandorresteijn@uva.nl)

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group  
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

advantages and disadvantages of different ruling systems have been discussed widely (e.g. Mehrens and Phillips 1989; McBee, Peters, and Waterman 2014). Yet, we have detected no consensus about the desirability of those advantages and disadvantages, nor about the possible consequences for the validity of the ultimate pass or fail decisions. Commonly, these discussions focus on one of three aspects of testing: (1) the *reliability* of examination grades, (2) how assessment policies affect *student behavior*; and (3) the *economic* costs of resits.

Concerning the *reliability* of examination grades, researchers often refer to classical test theory (Novick 1966), which postulates that each observed score consists of a true score and a measurement error. Examinations with a lower reliability contain larger measurement errors and therefore have, on average, a larger discrepancy between observed and true grades. The average grade is deemed more reliable as measurements errors cancel out when multiple assessments are used. Proponents of a compensatory rule have argued that this rule limits the number of students who unfairly fail an examination (Van Rijn, Beguin, and Verstralen 2012) and may therefore be considered most fair from a student perspective (Evers et al. 2010; Hambleton, Pitoniak, and Copella 2011; Möltner, Timbil, and Jünger 2015). This is not necessarily congruent with the perspective of the educational institute, however. Proponents of a conjunctive rule point toward the number of students who unfairly pass an examination under a compensatory rule, and thus that a conjunctive rule is better able to safeguard educational standards (Haladyna and Hess 1999).

Turning to *student behavior*, as pointed out in the literature, students (may) show different learning strategies, depending on the ruling system (Smits, Kelderman, and Hoeksma 2015; Kickert et al. 2019; Schmidt et al. 2021). If students were to strategically prepare for examinations - that is, when students selectively allocate study time based on the expected marginal returns - a compensatory rule may result in students developing knowledge hiatuses (Smits, Kelderman, and Hoeksma 2015). Students might neglect certain course elements once it is ascertained that they already passed the course.

Lastly, resits are costly from an *economic* perspective. For students, resits are costly because they are burdened with additional preparation and potential delay. For institutions, resits are costly because they have to organize additional assessment opportunities. This implies that if resits offer little to no learning gains, they might only incur unnecessary costs.

Reviewing the literature on this discussion, we encountered two studies in *Assessment & Evaluation in Higher Education* (Proud 2015; Arnold 2017) that evaluated whether a resit improves students' proficiency, assuming that if such an improvement were not found, this would show the redundancy of the conjunctive rule (and thus the pertinence of the compensatory rule). Using a *regression discontinuity design* (RDD; Thistlethwaite and Campbell 1960), both Proud (2015) and Arnold (2017) found little or no evidence for positive effects of resits on future performance, although Proud noted that a resit experience might (indirectly) benefit some students as the resit might change the way students prepare for subsequent examinations. This absence of evidence of a resit effect raises the question whether the educational gains of resits outweigh their costs, implicitly suggesting that a compensatory system might be preferred over a conjunctive system with resits.

The present aim is to point out that in Proud (2015) and Arnold (2017) the absence of evidence should not be mistaken for evidence of absence (Altman and Bland 1995; Parkhurst 2001). We argue that when trying to prove the absence of an effect, classical null-hypothesis testing is not always appropriate. Sufficient statistical power is necessary for such a claim. Yet, generally speaking, the power of *regression discontinuity analysis* (RDA) is low (e.g. Goldberger 1972). RDA requires a much larger sample size compared to the analysis in a randomized controlled trial (RCT) to achieve equivalent statistical accuracy, with estimates ranging from three to four times as large (Schochet 2009) up to nine to seventeen times as large (Deke and Dragoset 2012). Thus, it might very well be that effects of the ruling policy exist (as also suggested by Kickert et al. 2021; Yocarini et al. 2020; Schmidt et al. 2021), but had not been found

due to methodological shortcomings. With respect to this issue, Proud (2015) touched upon the issue of low power, albeit only implicitly.

To address the power issue, and hence to substantiate our claim that conclusion validity is at stake, we conducted an illustrative series of simulations. These show that both under realistic and relatively extreme conditions, RDA is ill-suited to detect a learning effect of a resit. Before turning to these simulations, we first briefly explain RDA and its associated research design.

### ***Regression discontinuity design (RDD) and regression discontinuity analysis (RDA)***

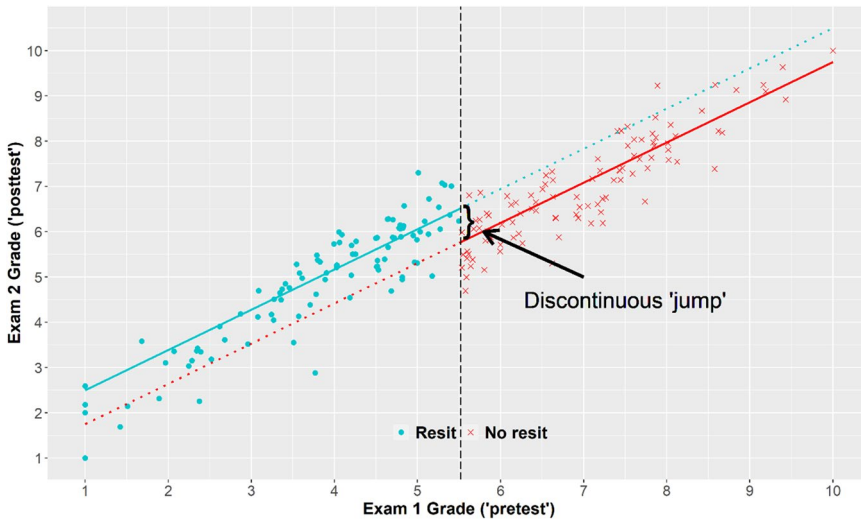
Regression discontinuity designs and regression discontinuity analyses have been widely used in educational research. Although RDD gained popularity in the economic domain (Van der Klaauw 2002), it was initially developed to evaluate the effectiveness of scholarship programs (Thistlethwaite and Campbell 1960). More recent examples include studies into the effect of class size on student performance (Angrist and Lavy 1999), financial aid on college attendance (Van der Klaauw 2002), and the effectiveness of a remedial summer school program (Matsudaira 2008). Before discussing the use of RDD in Proud (2015) and Arnold (2017), a short introduction of RDD is provided to inform readers who are unfamiliar with the method.

The regression discontinuity design can be considered as a multiple group pretest-posttest design. Contrary to an experimental design, group assignment is not random, but based on whether one's pretest score lies below or above a predetermined threshold (e.g. a cut-off score that suffices to pass a course). Individuals who have scored below a certain threshold are assumed to have undergone a different 'treatment' (e.g. are required to take a resit) than those who scored at or above the threshold. If the 'treatment' has an effect, one expects to see a discontinuous difference in the pretest-posttest scatterplot. This supposed discontinuity is the target in the subsequent (regression) analysis, hence the name of the analysis (regression discontinuity analysis; RDA) and its corresponding design (regression discontinuity design; RDD).

One of the major benefits of an RDD over an RCT is its convenient use. An RDD does not require an a priori random assignment to control and treatment groups because it is based upon post hoc assignment. Additionally, an RDD does not require needy participants to be assigned to a no-treatment group (Cappelleri and Trochim 2015). It has to be noted, however, that this advantage of RDD is also its weakness. The internal validity is lower for an RDD than for an RCT because RDD is merely a correlational design, making causal inferences problematic for an RDD because it is difficult to control for confounding variables. Also, the corresponding statistical analysis (RDA) has known weaknesses. As mentioned earlier, the power of an RDA is far lower than, for example, analyses of variance (ANOVAs), which are commonly used when analyzing RCTs.

To illustrate how a basic RDA can be conducted, imagine 200 students who take an examination (the pretest), where students who fail the examination are required to take a resit (the intervention). If the resit resulted in a substantial learning gain, the students taking a resit are expected to score higher on a subsequent examination (the posttest) than equivalent students who did not take a resit. Plotting the scores of the first examination ('pretest') against the scores on the second examination ('posttest'), as in Figure 1, is expected to reveal a 'discontinuous jump' at the pass/fail threshold – 5.5 on a 1 to 10 scale in this example – on the pretest measure.

In the most basic analysis, one regresses the posttest score on the pretest score and a dummy variable that indicates whether participants scored below (i.e. took a resit) or above (i.e. did not take a resit) the threshold (i.e. 5.5) as predictors. It is assumed that the lines are parallel but their intercepts differ. Next, the  $t$ -statistic of the dummy variable is used to test whether or not there is a significant discontinuous jump at the threshold. If there is, the conclusion would be that the discontinuity may have been caused by the treatment (while claiming that threats to internal validity, such as differences in history, have been refuted). If there is no discontinuity, however, the likelihood that the treatment did not have any consequences depends



**Figure 1.** Visualization of a 'discontinuous jump' using a regression discontinuity design.

on the established power, which is known to be generally weak. Therefore, RDA as a means to demonstrate that significant effects are absent is a weak method.

To demonstrate the inappropriateness of using RDA in the conjunctive versus compensatory rulings discussion we examined the following question: under what conditions, if any, does an RDA have sufficient power to detect a resit effect? We addressed this research question using a Monte Carlo study, which we will present next.

## Methodology

### Modelling the effect of a resit

Our simulations are theoretically grounded in classical test theory (Novick 1966) and are similar to earlier simulation studies on combining scores on multiple examinations (Douglas and Mislavy 2010; Van Rijn, Beguin, and Verstralen 2012; Yocarini et al. 2018). For a randomly chosen student, the observed score  $\bar{X}$  on an examination is modeled as:

$$X = \bar{X} + T + e \quad (1)$$

where  $T$  stands for the unobserved true score of the student,  $e$  for random measurement error, and  $\bar{X}$  is added for convenience to express  $T$  as a deviation from the group mean. Equation 1 is commonly generalized towards a population in which  $T$  varies among individuals, and it is assumed that both  $T$  and  $e$  are normally and independently distributed with a mean of zero. By definition, in classical test theory the following then holds:

$$\sigma_X^2 = \sigma_T^2 + \sigma_e^2 \quad (2)$$

The reliability  $\rho_{XX'}$  is defined as the ratio of the true score variance relative to the observed score variance:

$$\rho_{XX'} = \frac{\sigma_T^2}{\sigma_X^2} \quad (3)$$

In addition, the size of measurement error  $e$  can be expressed as:

$$\sigma_e = \sigma_x \sqrt{1 - \rho_{xx'}} \quad (4)$$

In the case of two examinations, each student has a pair of scores,  $X_1$  and  $X_2$ . The relation between these observed scores is modeled through  $\rho$ , the correlation between true scores  $T_1$  and  $T_2$ . Assuming means of zero and equal variances, the true score of the second examination,  $T_2$ , may be regressed on that of the first,  $T_1$ , using:

$$T_2 = \rho T_1 + \epsilon \quad (5)$$

where  $\epsilon$  is a prediction error with  $\sigma_\epsilon = \sigma_{T_2} \sqrt{1 - \rho^2}$ .

Although the effect of a resit has not been included in previous simulation studies (Van Rijn, Beguin, and Verstralen 2012; Yocarini et al. 2018), its implementation is rather straightforward. After failing the first examination (i.e.  $X_1$  is lower than cut-off score  $X^*$ ), the student prepares a resit which leads to an increase of  $\delta$  in her true score  $T_1$ , and its update may be written as:

$$T_1' = T_1 + \delta \quad (6)$$

where  $T_1$  is the original level of the true score on the first examination, and  $T_1'$  its final level.

After resitting the first examination the student takes the second examination. Equation 5 shows that an increase in  $T_1$  leads to an increase in  $T_2$ , and the effect of the resit on the observed score of the second examination is obtained by combining Equations 1, 5 and 6, giving:

$$X_2 = \bar{X}_2 + \rho(T_1 + \delta) + \epsilon + e \quad (7)$$

In this setup an observed score on the second examination is thus a function of the final level of the first true score, a prediction error and a measurement error, where the average grade of the second examination,  $\bar{X}_2$ , and the correlation between the first and second examination,  $\rho$ , are constants that determine the location and spread, respectively.

### Data generation

The data generation procedure used the grading scales and rules that are common in higher education in the Netherlands (cf. Arnold 2017). The simulated grades were thus expressed on a scale from 1 to 10 and rounded off to one decimal place; an examination was passed if the grade was equal to or larger than  $X^* = 5.5$ , and failed otherwise. In the first step, an observed score for the first examination  $X_1$  was obtained by drawing a true score for each simulee from a normal distribution ( $T_1 \sim N[0, \sigma_{T_1}^2]$ ) and adding both the average grade  $\bar{X}_1$  and a randomly drawn measurement error  $e_1$ . The standard deviations of the true scores and the measurement errors for the first examination were fixed to values that resulted in observed scores with a standard deviation of 1.5 and a reliability of 0.70, respectively, values often encountered in Dutch higher education (Smits, Kelderman, and Hoeksma 2015). In the second step, to include the effect of the resit, for each observation with an observed grade lower than the cut-score ( $X_1 < X^*$ ), a value of  $\delta$  was added to the original true score of the first examination (see Equation 6). Finally, Equation 7 was used to obtain an observed grade for the second examination. The standard deviations of the prediction errors and measurement errors for the second examination were fixed to values that, again, resulted in observed scores with a standard deviation of 1.5 and a reliability of 0.70. Both  $X_1$  and  $X_2$  were rounded off to one decimal place.

Proud (2015) and Arnold (2017) estimated a series of RDA models, some with and some without covariates. Because of varying effect sizes and statistical significance of the covariates in these models (and no resit effect was detected), it was unclear whether adding covariates would generally increase or decrease the statistical power to detect a resit effect, nor was it clear how to model the relation between covariates and other variables. It was therefore decided not to include covariates in the simulation, i.e. to stick to the basic application of RDA.

### Simulation design

In the simulation study the following four factors were varied: the sample size, the percentage of students taking a resit, the correlation between the two examinations, and the magnitude of the resit effect.

To study the effect of the *sample size*, three levels were used that represent common group sizes found in higher education settings and educational research. The lowest level, 100 observations, is a number that is often encountered in medium-sized courses in Dutch universities and equivalent to the sample size reported in Proud (2015). The highest level, 1,000 observations, was similar to the number of students in Arnold (2017). A sample size of 500 was chosen as an intermediate number.

To study the effect of the *proportion* of students taking a resit, two percentages were used: 25% and 50%. The percentage of resits was controlled through the average of the observed grades of the first examination. Under normality, using a cut-score of 5.5 and a standard deviation of 1.5, averages of 6.5 and 5.5 match with right-tail probabilities of 25% and 50%, respectively. Note that these percentages were approximate rather than exact (meaning that they varied a few percentage points from one draw to the next) as true scores were drawn from a population and because measurement errors were added to obtain observed scores. The average of the observed scores on the second examination  $\bar{X}_2$  was fixed at 6.0 throughout the simulations.

To study the effect of the *correlation between the examinations*, two correlation sizes were used to reflect the dependence between the two true scores ( $\rho$  in Equations 5 and 7): 0.30 and 0.50. In terms of Cohen's (1988) criteria for effect sizes, these values represent medium and high linear relationships, respectively.

To determine the effect of the *magnitude of the resit effect*, three effect sizes were used that reflect the gain in proficiency ( $\delta$  in Equations 6 and 7): 0, 0.50 and 0.80, scaled in units of  $\sigma_{T_1}$ . The first value represents the absence of a resit effect, and was added to evaluate type I errors. According to Cohen's (1988) guidelines, the latter two values represent medium and high gains, respectively.

The factors sample size, percentage of students taking a resit, correlation between examinations and resit effect produce a  $3 \times 2 \times 2 \times 3$  design. Within each cell of this design 100 replications were produced, yielding a total of 3,600 simulated data sets.

### Regression discontinuity analysis on generated datasets

In this simulation study the standard version of regression discontinuity analysis (RDA), as described by Proud (2015) and Arnold (2017), was applied. Each data matrix consisted of the scores on the regular sit of the first examination ( $X_1$ ), the scores on the second examination ( $X_2$ ), and a dummy variable  $D$  was created, which was set to one if  $X_1 \geq X^*$  (i.e. the student passed), and to zero otherwise. Using ordinary least squares the following regression equation was estimated:

$$E(X_2) = \beta_0 + \beta_1 D + \beta_2 (X_1 - X^*) \quad (8)$$



To evaluate the effect of a resit, a  $t$ -test was performed to test whether  $\beta_1$  deviated from 0. Similar to Proud (2015) and Arnold (2017) the hypothesis was tested against a two-sided alternative with a nominal probability of a type I error of 0.05.

### **Simulation outcomes and evaluation**

Two outcomes were evaluated: the power and the probability of a type I error. The primary outcome was the power to reject  $H_0 : \beta_1 = 0$  using RDA. Power estimates were obtained by calculating the proportion of instances that  $H_0$  was correctly rejected over the 100 replications in the design cells with a simulated resit effect (i.e.  $\delta > 0$ ). These outcomes were evaluated using the convention that a power of at least 0.80 represents a 'reasonable' value, as suggested by Cohen (1988, p. 56). As is common in power studies, the probability of a type I error was examined as a secondary outcome. This was done by comparing the proportion across 100 replications of incorrectly rejected  $H_0$  in the design cells in which a resit effect was absent (i.e.  $\delta = 0$ ) to the nominal probability of a type I error (i.e.  $\alpha = 0.05$ ). It was chosen to only fully report on substantial deviations from the nominal probability.

The simulations were conducted in *R* (R Core Team 2020, version 4.0.3). The *R* package *SimDesign* (Sigal and Chalmers 2016, version 2.1) was used to generate and analyze the data. The *R* syntax of the simulation is obtainable from the authors.

## **Results**

The Type I error rate was mostly close to the nominal value of 0.05, ranging from 0.03 to 0.09, and was unrelated to the design factors. It was therefore concluded that RDA showed valid Type I errors.

Turning to the results of the primary outcome, Table 1 shows the power estimates in the twenty-four cells of the design that included a resit effect. As can be expected, the power increased with an increasing sample size, an increasing correlation between the examinations, and an increasing effect size. The two levels of percentage of resits did not show a consistent power difference. Overall, the power to detect a resit effect was very low. A reasonable power was only encountered in the two cells with the largest sample size, a large correlation between the examinations, and a large resit effect: a power of 0.82 in the case of 50% resits and 0.91 in the case of 25% resits.

## **Discussion**

In line with the general critique on RDA (e.g. Schochet 2009), our simulation study showed that this type of analysis is expected to have very little power to detect the effect of a resit. If a resit effect were to exist in reality, using RDA is rather futile; effects would only be detected under improbable conditions, viz. when a very large sample is collected, the examinations are highly correlated, and the resit effect is large. Our results indicate that, due to its low power, RDA is an inappropriate method to make decisions about the appropriateness of assessment policies when using empirical data. In other words, the absence of significant effects in a sample should not be taken as evidence of absence of effects in the population. Therefore, the suggestion by Proud (2015) and Arnold (2017) that resits have little added value (and thus implicitly suggesting that a compensatory is to be favored over a conjunctive rule) has little conclusion validity.

It should be noted that although the power was reasonable in two combinations of the design factors, their usefulness will be limited in practice. First, even if the sample size

**Table 1.** Power of detecting change as a function of four design variables.

$N$	% resits	$\rho$	$\delta$	Power
100	25	0.3	0.5	0.07
100	50	0.3	0.5	0.05
100	25	0.5	0.5	0.10
100	50	0.5	0.5	0.07
100	25	0.3	0.8	0.04
100	50	0.3	0.8	0.09
100	25	0.5	0.8	0.18
100	50	0.5	0.8	0.16
500	25	0.3	0.5	0.12
500	50	0.3	0.5	0.12
500	25	0.5	0.5	0.28
500	50	0.5	0.5	0.23
500	25	0.3	0.8	0.22
500	50	0.3	0.8	0.30
500	25	0.5	0.8	0.48
500	50	0.5	0.8	0.56
1000	25	0.3	0.5	0.18
1000	50	0.3	0.5	0.23
1000	25	0.5	0.5	0.42
1000	50	0.5	0.5	0.42
1000	25	0.3	0.8	0.33
1000	50	0.3	0.8	0.37
1000	25	0.5	0.8	<b>0.91</b>
1000	50	0.5	0.8	<b>0.82</b>

Note.  $N$  is the number of students;  $\rho$  is the correlation between the two true scores; and  $\delta$  is the resit effect. Results with reasonable power ( $\geq 0.80$ ) are bold-faced.

requirement of at least 1,000 students is met, a number generally only met in very large courses, the power is only sufficient when both the examinations are highly correlated, and the resit effect is large, both of which would be very difficult to reach in practice. In this respect it is also worth noting that one may encounter a smaller proportion of students who are required to take a resit than tested in this study, which would make it even more difficult to detect a resit effect. These observations illustrate that our conclusions are universal and extend beyond the higher education context in the Netherlands. Second, these simulation results are valid in situations in which the model holds for all observations. To increase the sample size, and thus the power, one may consider merging data from different cohorts or similar courses. Yet, this leads to a situation in which each subset of students has its own true model, which, in all likelihood, will lead to less power to detect an overall ('pooled') resit effect. Aggregation of multiple samples should only be done if the statistical method can account for heterogeneous sample-specific effects (cf. Young 2001), which cannot be done through standard RDA.

Previous simulation studies also considered the impact of the reliability of the measures used for modeling (see, e.g. Trochim, Cappelleri, and Reichardt 1991). In the current study a constant value of 0.70 was used for the reliability of the simulated examinations. This value is typically encountered in Dutch higher education (Smits, Kelderman, and Hoeksma 2015) and as a rule of thumb represents 'adequate' measurement precision (Evers et al. 2013). However, there has been a call for higher standards given the poor quality of the grading in higher education (e.g. Bloxham et al. 2016). This raises the question whether the power to detect a resit effect would increase if examinations met higher reliability standards. To that end, a sensitivity analysis was performed, not shown herein, where instead of examinations with 'adequate' reliability, examinations with 'excellent' reliability (i.e. a value of 0.90; Evers et al. 2013) were used. Although the power increased somewhat in most cases, the pattern did not change; there were still only two cells in which a power of at least 0.80 was achieved. Based on this analysis, it may be concluded that a higher measurement precision does not solve the issue of RDD's low statistical power. Moreover, when the reliability of examinations is less than 'adequate', power is expected to be even lower.

This study focused on using RDA in the context of detecting a resit effect. The authors acknowledge that because of its intuitive and straightforward application, RDA has a great natural appeal to educational researchers for wider application, particularly as data like student grades are easily obtainable from university administration offices. For example, when pass or fail decisions are based on the average of multiple assessments, one could wonder whether students show strategic behavior when preparing for these examinations. In a compensatory system, strategic students who are certain to pass may be inclined to put less effort in their preparation for subsequent assessments and are therefore expected to score lower on the final assessment(s) than to be expected based on their true score. To test whether, and to which extent, students exhibit such behavior, it would be tempting to apply RDA in this context. Yet, like before, it would be an inappropriate approach. It would be inappropriate not only due to a lack of power, but also because of the discrepancy between what strategic behavior implies and the assumptions of an RDA; if only some of the students would show strategic behavior, a mix of positive and negative relations between examinations would be expected. Rather than a jump near a threshold (as is assumed by RDA), strategic behavior would imply a mixture of multiple non-parallel regression lines across the grade spectrum. A small simulation study similar to the current one, not shown herein, where strategic behavior was assumed, indeed showed that RDA had very low power (mostly  $<0.10$ ) to detect such behavior under all conditions.

To conclude, when choosing between a compensatory or conjunctive rule (or any comparison between assessment policies for that matter), it is important to examine the reliability and validity of these assessment policies. Given that RDD has low power, other methods have to be used in order to adequately compare assessment policies. Moreover, rather than only focusing on the issue of which statistical method to use, to truly assess how assessment policies might affect student performance requires researchers to use sophisticated study designs. In this respect, rather than post hoc grade evaluation, it may be more appropriate to conduct experimental studies, or perhaps the most insightful approach may be to simply ask students how they would react to different assessment policies. Irrespective of the method it is important to be aware of the aphorism that absence of evidence is not evidence of absence.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## ORCID

Chevy van Dorresteijn  <http://orcid.org/0000-0002-6240-185X>

Kees-Jan Kan  <http://orcid.org/0000-0003-0088-9906>

Niels Smits  <http://orcid.org/0000-0003-3669-9266>

## References

- Altman, D. G., and J. M. Bland. 1995. "Statistics Notes: Absence of Evidence is Not Evidence of Absence." *BMJ (Clinical Research ed.)* 311 (7003): 485. doi:10.1136/bmj.311.7003.485.
- Angrist, J. D., and V. Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement." *The Quarterly Journal of Economics* 114 (2): 533–575. doi:10.1162/003355399556061.
- Arnold, I. 2017. "Resitting or Compensating a Failed Examination: Does It Affect Subsequent Results?" *Assessment & Evaluation in Higher Education* 42 (7): 1103–1117. doi:10.1080/02602938.2016.1233520.
- Biggs, J., and C. Tang. 2007. *Teaching for Quality Learning at University Maidenhead*. 3rd ed. Maidenhead: Hill Education.

- Bloxham, S., B. den-Outer, J. Hudson, and M. Price. 2016. "Let's Stop the Pretence of Consistent Marking: Exploring the Multiple Limitations of Assessment Criteria." *Assessment & Evaluation in Higher Education* 41 (3): 466–481. doi:10.1080/02602938.2015.1024607.
- Cappelleri, J. C., and W. M. Trochim. 2015. "Regression Discontinuity Design." In *International Encyclopedia of the Social & Behavioral Sciences*, edited by J. D. Wright. 2nd ed., 152–159. Amsterdam: Elsevier. doi:10.1016/B978-0-08-097086-8.44049-3.
- Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Deke, J., and L. Dragoset. 2012. *Statistical power for regression discontinuity designs in education: Empirical estimates of design effects relative to randomized controlled trials*. Working paper. Mathematica Policy Research. <https://eric.ed.gov/?id=ED533141>
- Douglas, K. M., and R. J. Mislevy. 2010. "Estimating Classification Accuracy for Complex Decision Rules Based on Multiple Scores." *Journal of Educational and Behavioral Statistics* 35 (3): 280–306. doi:10.3102/1076998609346969.
- Evers, A., Hagemester, C. Höstmaelingen, A. Lindley, P. Muñiz, and J. Sjöberg. 2013. *EFPA Review Model for the Description and Evaluation of Psychological and Educational Tests: Test Review Form and Notes for Reviewers Version 4.2.6*. Brussels: European Federation of Psychologists' Associations.
- Evers, A., K. Sijtsma, W. Lucassen, and R. R. Meijer. 2010. "The Dutch Review Process for Evaluating the Quality of Psychological Tests: History, Procedure, and Results." *International Journal of Testing* 10 (4): 295–317. doi:10.1080/15305058.2010.518325.
- Goldberger, A. S. 1972. *Selection Bias in Evaluating Treatment Effects: The Case of Interaction*. Madison, WI: Institute for Research on Poverty, University of Wisconsin.
- Haladyna, T., and R. Hess. 1999. "An Evaluation of Conjunctive and Compensatory Standard-Setting Strategies for Test Decisions." *Educational Assessment* 6 (2): 129–153. doi:10.1207/S15326977EA0602\_03.
- Hambleton, R. K., M. J. Pitoniak, and J. M. Copella. 2011. "Essential Steps in Setting Performance Standards on Educational Tests and Strategies for Assessing the Reliability of Results." In *Setting Performance Standards: Foundations, Methods, and Innovations*, edited by G. J. Cizek, 47–76. 2nd ed. New York, NY: Routledge.
- Kickert, R., M. Meeuwisse, K. M. Stegers-Jager, G. V. Koppenol-Gonzalez, L. R. Arends, and P. Prinzie. 2019. "Assessment Policies and Academic Performance within a Single Course: The Role of Motivation and Self-Regulation." *Assessment & Evaluation in Higher Education* 44 (8): 1177–1190. doi:10.1080/02602938.2019.1580674.
- Kickert, R., M. Meeuwisse, L. R. Arends, P. Prinzie, and K. M. Stegers-Jager. 2021. "Assessment Policies and Academic Progress: Differences in Performance and Selection for Progress." *Assessment & Evaluation in Higher Education* 46 (7): 1140. doi:10.1080/02602938.2020.1845607.
- Matsudaira, J. D. 2008. "Mandatory Summer School and Student Achievement." *Journal of Econometrics* 142 (2): 829–850. doi:10.1016/j.jeconom.2007.05.015.
- McBee, M. T., S. J. Peters, and C. Waterman. 2014. "Combining Scores in Multiple-Criteria Assessment Systems: The Impact of Combination Rule." *Gifted Child Quarterly* 58 (1): 69–89. doi:10.1177/0016986213513794.
- Mehrens, W. A., and S. E. Phillips. 1989. "Using College GPA and Test Scores in Teacher Licensure Decisions: Conjunctive versus Compensatory Models." *Applied Measurement in Education* 2 (4): 277–288. doi:10.1207/s15324818ame0204\_1.
- Möltner, A., S. Timbil, and J. Jünger. 2015. "The Reliability of the Pass/Fail Decision for Assessments Comprised of Multiple Components." *GMS Zeitschrift Für Medizinische Ausbildung* 32 (4). doi:10.3205/zma000984.
- Novick, M. R. 1966. "The Axioms and Principal Results of Classical Test Theory." *Journal of Mathematical Psychology* 3 (1): 1–18. doi:10.1016/0022-2496(66)90002-2.
- Parkhurst, D. F. 2001. "Statistical Significance Tests: Equivalence and Reverse Tests Should Reduce Misinterpretation: Equivalence Tests Improve the Logic of Significance Testing When Demonstrating Similarity is Important, and Reverse Tests Can Help Show That Failure to Reject a Null Hypothesis Does Not Support That Hypothesis." *Bioscience* 51 (12): 1051–1057. doi:10.1641/0006-3568(2001)051[1051:SSTEAR]2.0.CO;2.
- Prigoff, J., M. Hunter, and R. Nowygrad. 2021. "Medical Student Assessment in the Time of COVID-19." *Journal of Surgical Education* 78 (2): 370–374. doi:10.1016/j.jsurg.2020.07.040.

- Proud, S. 2015. "Resits in Higher Education: Merely a Bar to Jump over, or Do They Give a Pedagogical 'Leg Up'?" *Assessment & Evaluation in Higher Education* 40 (5): 681–697. doi:10.1080/02602938.2014.947241.
- R Core Team. 2020. *RStudio: Integrated development for R*. RStudio, inc. Boston, MA. <http://www.rstudio.com/>
- Schmidt, H. G., G. J. Baars, P. Hermus, H. T. van der Molen, I. J. Arnold, and G. Smeets. 2021. "Changes in Examination Practices Reduce Procrastination in University Students." *European Journal of Higher Education*. doi:10.1080/21568235.2021.1875857.
- Schochet, P. Z. 2009. "Statistical Power for Regression Discontinuity Designs in Education Evaluations." *Journal of Educational and Behavioral Statistics* 34 (2): 238–266. doi:10.3102/1076998609332748.
- Sigal, M. J., and R. P. Chalmers. 2016. "Play It Again: Teaching Statistics with Monte Carlo Simulation." *Journal of Statistics Education* 24 (3): 136–156. doi:10.1080/10691898.2016.1246953.
- Smits, N., H. Kelderman, and J. B. Hoeksma. 2015. "Een Vergelijking Van Compensatoir en Conjunctief Toetsen in Het Hoger Onderwijs [A comparison of compensatory and conjunctive testing in higher education]." *Pedagogische Studiën* 92 (4): 150–160.
- Tan, C. K., W. L. Chua, C. K. F. Vu, and J. P. E. Chang. 2021. "High-Stakes Examinations during the COVID-19 Pandemic: To Proceed or Not to Proceed, That is the Question." *Postgraduate Medical Journal* 97 (1149): 427–431. doi:10.1136/postgradmedj-2020-139241.
- Thistlethwaite, D. L., and D. T. Campbell. 1960. "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment." *Journal of Educational Psychology* 51 (6): 309–317. doi:10.1037/h0044319.
- Trochim, W. M. K., J. C. Cappelleri, and C. S. Reichardt. 1991. "Random Measurement Error Does Not Bias the Treatment Effect Estimate in the Regression-Discontinuity Design: II when an Interaction Effect is Present." *Evaluation Review* 15 (5): 571–604. doi:10.1177/0193841X9101500504.
- Van der Klaauw, W. 2002. "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach." *International Economic Review* 43 (4): 1249–1287. doi:10.1111/1468-2354.t01-1-00055.
- Van Rijn, P. W., A. A. Beguin, and H. Verstralen. 2012. "Educational Measurement Issues and Implications of High Stakes Decision Making in Final Examinations in Secondary Education in The Netherlands." *Assessment in Education: Principles, Policy & Practice* 19 (1): 117–136. doi:10.1080/0969594X.2011.591289.
- Walstad, W. B. 2001. "Improving Assessment in University Economics." *The Journal of Economic Education* 32 (3): 281–294. doi:10.1080/00220480109596109.
- Yocarini, I. E., S. Bouwmeester, G. Smeets, and L. R. Arends. 2018. "Systematic Comparison of Decision Accuracy of Complex Compensatory Decision Rules Combining Multiple Tests in a Higher Education Context." *Educational Measurement: Issues and Practice* 37 (3): 24–39. doi:10.1111/emip.12186.
- Yocarini, I. E., S. Bouwmeester, G. Smeets, and L. R. Arends. 2020. "Allowing Course Compensation in Higher Education: A Latent Class Regression Analysis to Evaluate Performance on a Follow-up Course." *Assessment & Evaluation in Higher Education* 45 (5): 728–740. doi:10.1080/02602938.2019.1693494.
- Young, J. W. 2001. *Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis*. Research report no. 2001-6. College Entrance Examination Board. <https://eric.ed.gov/?id=ED562661>