International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 9, Issue.2, March-April 2026, page no. 131-137

To cite this article: Jose A. Martínez (2026). ASSESSING DIFFERENTIAL DIFFICULTY IN PARALLEL ASSESSMENT SLOTS: A MISSPECIFICATION-BASED STATISTICAL APPROACH TO GRADING EQUITY, International Journal of Education and Social Science Research (IJESSR) 9 (2): 131-137 Article No. 1210, Sub Id 1867

ASSESSING DIFFERENTIAL DIFFICULTY IN PARALLEL ASSESSMENT SLOTS: A MISSPECIFICATION-BASED STATISTICAL APPROACH TO GRADING EQUITY

Jose A. Martínez

Department of Business Economics, Technical University of Cartagena – Member of European University of Technology,
Calle Real, 3, 30201, Cartagena, Spain

DOI : https://doi.org/10.37500/IJESSR.2026.9209

ABSTRACT

Scheduling assessment activities across multiple time slots is an increasingly common institutional response to student diversity, particularly the growing proportion of working students in higher education. However, when distinct question sets are used across slots to prevent information leakage, differential difficulty may emerge, inadvertently penalizing students assigned to a harder version. This paper presents a practical case study from the course Marketing Management II (second year of the Bachelor in Business Administration and Management, Technical University of Cartagena, Spain), where an online formative quiz was administered in two non-overlapping time slots with different question sets. Building on Spanos's (2019) misspecification testing philosophy, a two-sample t-test was applied after systematically validating its underlying assumptions through runs tests, t-plots, Levene's test for equality of variances, and the Shapiro–Wilk normality test. Results confirmed a statistically significant mean difference of 2.29 points favouring the morning group ($T = 5.24$, $df = 124$, $p < .001$). As a corrective measure, 1.2 points were added to the scores of the afternoon group, substantially reducing the performance gap. The case illustrates both the feasibility and the importance of applying rigorous inferential tools to detect and correct unfair grading differentials in educational assessment.

KEYWORDS: assessment equity; grading fairness; misspecification testing; two-sample t-test; higher education

1. INTRODUCTION

The right of every student to be assessed fairly, objectively, and under conditions that afford an equal opportunity to demonstrate their knowledge and competences is a foundational principle of educational quality. Assessment equity requires that score differences between students reflect genuine differences in learning outcomes, rather than artefacts of the evaluation instrument or the conditions under which it is administered [1,2]. When assessment procedures introduce systematic advantages or disadvantages that are unrelated to student ability -such as differences in item difficulty, administration

conditions, or available time- the validity of the resulting scores is compromised, and the grades awarded may fail to fulfil their core function as fair measures of academic attainment.

This concern is particularly acute when assessments are administered across multiple time slots with different question sets. The use of parallel forms is a widely adopted strategy in higher education to accommodate students with diverse scheduling constraints -such as those who combine study with part-time or full-time employment- while simultaneously protecting the integrity of the evaluation by preventing information leakage between cohorts. However, parallel forms are difficult to make psychometrically equivalent in practice, creating a situation in which students are nominally evaluated on the same learning objectives but under substantively unequal conditions [3]. Left undetected and uncorrected, such disparities constitute a source of systematic bias that violates the principle of equal treatment and undermines institutional commitments to assessment validity and fairness.

Despite the normative centrality of assessment equity in quality assurance frameworks for higher education, the routine statistical verification of parallel form equivalence remains uncommon in everyday university teaching practice [4]. In most cases, the comparability of parallel assessment forms is assumed on the basis of instructor judgement rather than demonstrated through empirical analysis. This gap between normative aspiration and practical implementation motivates the present work.

This paper reports a concrete case study arising from the course Marketing Management II, a compulsory second-year subject of the Bachelor in Business Administration and Management (GADE) at the Technical University of Cartagena (UPCT), Spain. An online formative quiz was offered in two non-overlapping time slots with different question sets. Post hoc statistical analysis revealed a substantial and statistically significant difference in mean scores, raising a serious fairness concern that required an evidence-based corrective response. The analytical framework employed draws on the misspecification testing methodology of Aris Spanos [5,6], which requires the systematic validation of all statistical model assumptions prior to inference, thereby ensuring that the resulting conclusions rest on solid probabilistic foundations.

The remainder of the paper is structured as follows. Section 2 describes the assessment context and the statistical methodology. Section 3 presents the results. Section 4 discusses the implications of the findings, and Section 5 offers conclusions and recommendations for practice.

2. METHODOLOGY

2.1 Assessment Context

Marketing Management II is a core second-year course within the GADE programme at UPCT, covering marketing strategy, consumer behaviour, and commercial decision-making. In the academic year under study (2024–2025), enrolled students were distributed across three classroom groups (A, B, and C). The formative quiz activity analysed here was administered through the institutional virtual learning environment (VLE) and comprised six multiple-choice questions of a theoretical–practical

nature, each offering four possible answers, of which only one was correct. Incorrect answers carried a penalty of one-third of the value of a correct answer, and the total score was expressed on a minimum-10 scale.

In order to accommodate students with work commitments, family issues or other personal circumstances who might face difficulties attending a fixed assessment slot, the activity was offered in two non-overlapping sessions. The morning slot was available between 08:00 and 10:00, providing a two-hour window during which students could begin the quiz at any point of their choosing; the maximum permitted completion time was 40 minutes. The afternoon slot operated under identical conditions, a two-hour window from 20:00 to 22:00, with a maximum of 40 minutes, but featured an entirely different set of six questions. Students self-selected into the session of their choice. The deliberate use of different question sets across the two slots was intended to prevent students from the morning session sharing item content with those in the afternoon. This precautionary measure, while educationally justified, created the possibility of unequal item difficulty between the two forms, and therefore of unequal conditions of assessment.

A total of 126 students completed the activity: 69 in the morning session and 57 in the afternoon session. Because the activity contributed to the overall course grade, the identification and correction of any differential difficulty was both ethically obligatory and practically consequential for student welfare.

2.2 Statistical Model

The statistical question of interest is whether group membership (morning vs. afternoon slot) is systematically associated with differences in quiz scores. This is formalised as a simple linear regression model:

$$Y_t = \beta_0 + \beta_1 x_t + u_t, \qquad t \in \mathbb{N}$$

where Y_t denotes the quiz score of student t; x_t is a binary indicator equal to 0 for students in the morning session and 1 for those in the afternoon session; β_0 represents the expected mean score of the morning group; β_1 captures the mean difference between the afternoon and morning groups (the parameter of inferential interest); and u_t is a stochastic error term where $Cov(u|x_t)$=0. Under this parameterisation, the standard two-sample t-test for equality of means is algebraically equivalent to the t-test of $H_o: \beta_1 = 0$ in the above regression model. The model assumes that the error terms are independent, identically distributed (IID), and follow a normal distribution, assumptions that must be empirically validated prior to inference.

2.3 Misspecification Tests

Following Spanos's misspecification (M-S) testing framework [5,6], the probabilistic assumptions of the model were verified through a systematic battery of diagnostic tests applied to each group

International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 9, Issue.2, March-April 2026, p no. 131-137

separately before any inferential conclusion was drawn. This three-stage approach -(i) model specification, (ii) assumption validation, (iii) conditional inference- ensures that the statistical conclusions are not artefacts of model misspecification, a discipline that is especially important in educational research where sample sizes are modest and conclusions have direct consequences for student welfare.

Independence and identical distribution (IID) were assessed through a runs test [7] applied to the sequence of scores ordered by submission time, complemented by visual inspection of t-plots (scatter plots of scores against submission order). For the morning group, the runs test yielded for the morning group: $z = -1.56; p = 0.12; n = 69$. For the afternoon group: $z = -1.98; p = 0.05; n = 57$. In neither case was the IID null hypothesis rejected, and no temporal pattern was observable in the t-plots (Figure 1).

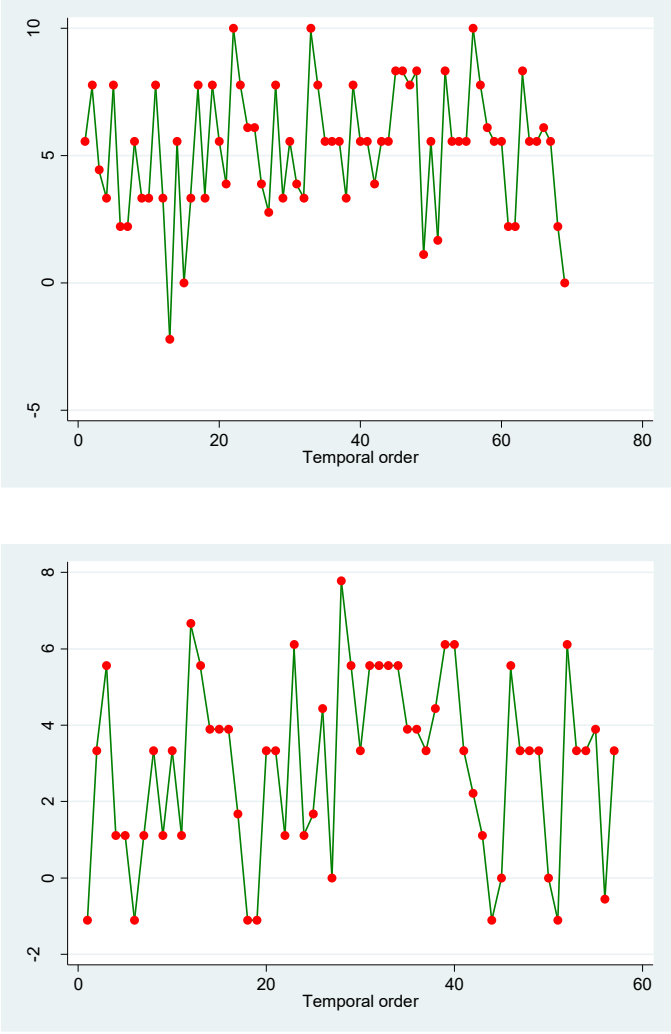

Figure 1. t-plots for the scores of both groups

International Journal of Education and Social Science Research

ISSN 2581-5148

Vol. 9, Issue.2, March-April 2026, p no. 131-137

Equality of variances across groups was examined using an F-ratio test, which yielded: $f = 0.89; p = 0.66$, providing no evidence against homoscedasticity. Normality within each group was assessed with the Shapiro–Wilk test [8], yielding $W = 0.98; p = 0.268$ for the morning group and $W = 0.962; p = 0.069$ for the afternoon group; normality was not rejected in either case. With all probabilistic assumptions confirmed, the two-sample t-test was deemed a statistically legitimate inferential tool for these data.

3. RESULTS

The morning group achieved a mean score of 5.26 (SD = 2.50; 95% CI: [4.66, 5.86]), while the afternoon group achieved a mean of 2.97 (SD = 2.36; 95% CI: [2.35, 3.60]). The non-overlapping confidence intervals already provided strong visual evidence of a substantial between-group difference.

A two-sample t-test with equal variances (confirmed by the preceding F-test) was conducted and yielded $T = 5.24; p < 0.001$, with an estimated mean difference of 2.29 points (99% CI: [1.15, 3.43]). This finding constitutes strong statistical evidence that the afternoon group's systematically lower performance is incompatible with the null hypothesis of equivalent form difficulty and must therefore be attributed -at least in substantial part- to greater item difficulty in the afternoon quiz.

Given this result, an upward score adjustment of 1.2 points was applied to all students in the afternoon session as a corrective measure aimed at restoring assessment equity. Under the assumption of equal variance, this brought the adjusted afternoon mean to 4.17. A follow-up t-test on the adjusted scores yielded $T = 2.49; p = 0.014 > 0.01$, with a residual mean difference of 1.09 points (99% CI: [-0.05, 2.23]). The correction substantially reduced the gap, making the difference of both means no significant at 99%.

4. DISCUSSION

The results of this case study carry implications at three levels: methodological, pedagogical, and institutional.

At the methodological level, the application of Spanos's misspecification testing framework proved essential to ensuring the legitimacy of the inferential conclusions. The systematic pre-inference validation of the IID, homoscedasticity, and normality assumptions transformed the t-test from a routine computation into a statistically defensible procedure whose conclusions can withstand scrutiny. This discipline -often bypassed in everyday educational research- should be regarded as a minimum standard whenever statistical tests are used to inform consequential decisions about student grades.

At the pedagogical level, the case illustrates that even well-intentioned assessment designs can

inadvertently introduce inequities. The decision to offer two time slots was motivated by a genuine concern for inclusivity, particularly for working students. Yet the unavoidable use of different question sets created a structural vulnerability that instructor judgement alone was insufficient to eliminate.

At the institutional level, the corrective action taken, adding 1.2 points to the afternoon group's scores, illustrates what an evidence-based, transparent, and documentable response to an identified fairness violation looks like in practice. The correction was not arbitrary: it was derived directly from the inferential output of a formally validated statistical model, providing a principled and reproducible rationale. Institutions should consider establishing formal protocols for this type of post hoc verification.

The main limitation of this study is that a possible alternative explanation is possible if the systematic differences between groups were driven by any other variables (e.g. better students vs worse students). However, observational inspection of database in the VLE did not detect any systematic difference in choosing the slots among students who attended regularly to class or who presented singular motivations.

5. CONCLUSION

This paper has demonstrated that statistical thinking can be applied to routine grading decisions in higher education. The case of Marketing Management II shows that a simple, well-validated two-sample t-test is sufficient to detect substantial and statistically significant differential difficulty between parallel assessment forms, and that such detection provides the evidential basis for a principled corrective response.

Three practical recommendations emerge from this case. First, any assessment administered across multiple time slots with different question sets should be subjected to a formal post hoc equivalence check before final grades are recorded. Second, the statistical model underlying such a check should be validated through misspecification tests before inference is drawn. Third, any corrective adjustment should be derived from and justified by the inferential output of the validated model, rather than relying on intuition or ad hoc decisions.

These recommendations are feasible within the resource constraints of ordinary university teaching and require no specialised psychometric infrastructure. Their adoption would represent a meaningful step towards operationalising the commitment to assessment equity that lies at the heart of quality assurance in higher education.

REFERENCES

[1] AERA, APA, & NCME. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

[2] Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), Educational Measurement (4th ed., pp. 221–256). Westport, CT: American Council on Education / Praeger.

[3] Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and Validating Test Items*. New York: Routledge.

[4] Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, *4*(3), 365–379.

[5] Spanos, A. (2018). Mis-Specification Testing in Retrospect, J*ournal of Economic Surveys, 32*, 541—577.

[6] Spanos, A. (2019). *Probability theory and statistical inference: Empirical modeling with observational data* (2nd ed.). Cambridge University Press.

[7] Wald, A., & Wolfowitz, J. (1940). On a test whether two samples are from the same population. Annals of Mathematical Statistics, 11(2), 147–162.

[8] Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3–4), 591–611.

Author Profile

Jose A. Martínez is an Associate Professor in the Department of Business Economics at the Universidad Politécnica de Cartagena (UPCT), Spain. His research interests encompass environmental economics, public health policy, and the intersections between business management and sustainability.