

A Minimax Optimization Model for Virtual Reality Experiential Learning

Xucheng Zhou¹, Sirirat Petsangsri^{2*}, Thanin Ratanaolarn³

¹School of Industrial Education and Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok, 10520, Thailand. E-mail: zjyc1heng@163.com, ORCID: <https://orcid.org/0009-0004-3928-1644>

²School of Industrial Education and Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok, 10520, Thailand. E-mail: sirirat.pe@kmitl.ac.th, ORCID: <https://orcid.org/0000-0002-4828-1740>

³School of Industrial Education and Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok, 10520, Thailand. E-mail: thanin.ra@kmitl.ac.th, ORCID: <https://orcid.org/0000-0003-1612-4852>

*Corresponding Author: Sirirat Petsangsri, E-mail: sirirat.pe@kmitl.ac.th

Abstract

This study develops a minimax optimization model for evaluating virtual reality experiential learning when an intervention is expected to improve several educational outcomes at the same time. Conventional evaluations usually test each outcome separately or combine outcomes through arbitrary weights. Both approaches can conceal a weak result in one domain. The proposed model instead selects the instructional strategy that maximizes the worst standardized lower-bound gain across outcomes. An empirical demonstration uses a supplied pretest-posttest dataset containing 100 observations, two groups, and composite scores for cultural competence and emotional intelligence. The data contain no missing values. The baseline profiles of the two groups are pairwise identical, so the analysis combines heteroskedasticity-robust analysis of covariance with pair-based bootstrap and sensitivity procedures. Adjusted posttest differences favor the virtual reality group for cultural competence (0.602, 95% CI [0.522, 0.683]) and emotional intelligence (0.600, 95% CI [0.527, 0.673]). After baseline-standardization, the estimated gains are 1.271 and 1.350 standard deviations. The conservative minimax lower-bound score is 1.087, with cultural competence defining the binding outcome. The same strategy is selected under change-score, paired, rank-based, trimmed-mean, and leave-one-pair-out analyses. The model offers a transparent alternative to weighted composite evaluation while preserving outcome balance and uncertainty for instructional decisions.

Keywords: minimax optimization; virtual reality; experiential learning; cultural competence; emotional intelligence; robust decision-making

1. Introduction

VR has transitioned from lab tech to a real-world teaching tool in higher education. Its educational value is evident. In a virtual setting, learners could be transported into an environment that is hard to recreate in a traditional classroom, such as a spatial, cultural, scientific or professional environment. It can also facilitate repeated practice, controlled exposure, and immediate interaction. The wider movement towards digitally mediated higher education has gained momentum following the unprecedented restructuring of university teaching over the course of the pandemic (Bond et al., 2021; Rapanta et al., 2020). But the experience is by no means uniformly positive. A recent review revealed that the learning outcomes of immersive systems are influenced by the guidance provided, task design, cognitive demands, and the congruence between the technology and the learning outcome (Di Natale et al., 2020), as well as presence and engagement (Hamilton et al., 2021; Radianti et al., 2020; Wu et al., 2020). It is then a question of whether virtual reality is attractive. It is whether a particular design can enhance outcomes and whether it does so in a balanced way.

This assessment challenge becomes a challenge if an intervention is intended to have an impact on more than one educational domain. A virtual reality experiential-learning program can be designed to enhance knowledge, cultural competence, emotional intelligence, motivation, transfer, and collaboration. Typically, each endpoint is tested individually. Another common method is to use weighted endpoints, which is done by averaging the endpoints. The conclusion may be broken up by separate tests. A weak domain can be masked with a strong domain if a sufficient amount of weight is given to the strong domain. Both approaches are particularly problematic if the intervention is justified as an integrated pedagogy. A programme that has a clearly identified learning aim and a strong impact on one outcome but not the other may not meet the programme's educational purpose.

The solution to this is provided by minimax optimization. A maximin rule chooses the strategy that has the highest performance guaranteed under the worst of conditions. The concept of robust decision making is well known. A decision is compared with negative conditions/uncertain conditions and the chosen alternative will be the one that has the best worst possible outcome (Rahimian & Mehrotra, 2022; Shapiro, 2021). The uncertain states may be seen as outcome domains in the educational context. Then the decision-maker asks himself this simple question: based on this estimation uncertainty, which instructional strategy yields the greatest minimum gain over all outcomes required? This formulation eliminates the need to come up with preference weights. It also lets you see the outcome of the binding. This paper proposes a general minimax model for multi-outcome experiential learning and illustrates it using a pretest - posttest dataset provided by the author. The workbook has half of the records set aside for one half and the other half for the other half. There are four composite scores: cultural competence pre and post-instruction; and emotional intelligence pre and post-instruction. The following study narrative outlines the virtual reality experiential-learning condition and comparison condition. The groups are numbered in the workbook. Group 2 is considered the EL-VR condition and group 1 is the comparison condition, as described in the provided narrative and observed intervention pattern. The data owner is responsible for confirming this mapping before submitting.

The data that can be obtained is limited and useful. They include full-scale scores, but do not provide item responses, demographic controls, intervention dose, session-level process measures, institutional identifiers, or calendar dates. The two groups also have the same baseline profiles in each of the 50 cross-group pairs. This pattern represents either an exact match, duplicate baseline records or a pre-processing operation. No source materials provide documentation of either the correct explanation. The analysis doesn't hide the feature but rather puts it under the rubric of integrity and dependence. The main models employ ANOVA that is robust to heteroskedasticity. Sensitivity checks include paired-based bootstrap, paired-clustered inference, paired comparisons, rank tests, trimmed means and leave-one-pair-out calculations.

The article has four contributions. The first is that it constructs a scale-adjusted maximin model which can compare instructional strategies that are measured on different scales. Second, it captures the uncertainty of sampling in the decision-making process, not just to provide lower-bound estimates but to minimize them. Thirdly, it distinguishes between the mathematical rule used to make decisions and the causal inferences made with the data. The model is general, and the empirical application is limited. Fourth, it offers a repeatable method of pinpointing the least well-protected educational outcome. This is particularly helpful as a binding result can help inform further iterations of instruction.

The study is guided by the following questions: RQ1 is a question that determines whether differences between the EL-VR and comparison group exist in both outcome measures and these differences are positive with pretest adjustment. RQ2 seeks to determine which strategy will be chosen if the decision criterion is to maximize the worst (standardized) lowerbound gain. RQ3: Does the chosen strategy stay the same for other statistical and normalization options? Figure 1 summarizes the analytical logic. The literature on education and optimization is then reviewed, the minimax model is defined and the empirical results are reported, followed by a discussion of the implications that can be drawn from the data provided.



Figure 1. Minimax decision architecture for outcome balancing.

2. Literature Review and Theoretical Basis

2.1 Virtual reality as an experiential learning environment

Immersive virtual reality can induce a feeling of being present in a space and can allow activities that are hard to accomplish using static media. In higher education, these affordances have been applied to different contexts such as laboratory, field trips, clinical training, heritage interpretation, engineering and professional simulation. Systematic reviews have revealed the field has grown rapidly, but not without some inconsistencies in research quality. Small sample size, brief interventions, minimal comparisons, and inadequate reporting continue to be prevalent (Radianti et al., 2020; Hamilton et al., 2021). The current literature indicates that HMDs have a beneficial effect on learning performance, but the results depend on the instructional design and characteristics of the studies (Kaplan et al., 2021; Villena-Taranilla et al., 2022; Wu et al., 2020).

The difference between being immersed in technology and pedagogically structured. The difference lies at the heart and center of technological immersion and pedagogical structure. Presence can distract attention, but can also be a mental drain. Novelty can make someone happy but not help them to have a lasting grasp on the material. This tension is explained by the Cognitive Affective Model of Immersive Learning which relates technological affordances, presence, agency, affective processing, cognitive load, and learning strategies (Makransky & Petersen, 2021). Experimental studies also suggest that the appropriate pairing of immersive experiences with generative activity, guidance, and reflection is beneficial to learning (Parong & Mayer, 2021; Petersen et al., 2020). A review of serious virtual environments says just that. Apart from visual richness, the learning task and the feedback and instructional sequence are more important (Checa & Bustillo, 2020).

The experiential learning pedagogical frame provides a useful pedagogical frame as it sees experience as food for reflection, conceptualization, and the next action. Recently, it has been suggested that concrete activity alone is not enough! Experience is only valuable to the learner when it is interpreted and transformed. (Morris, 2020). This process can be facilitated in a virtual environment when learners are presented with a meaningful situation, are involved in some action, talk about their experiences, and relate their experiences to concepts. They can also not work out if they are not linked to clear learning objectives. The reason is that immersion does not directly influence cultural competence and emotional intelligence, which is important for the current research. They need to be interpreted, to have a perspective, to be self-controlled and to have a social meaning.

2.2 Multi-domain educational outcomes

Learning is multi-dimensional and hence, the interventions in education have multiple outcomes. An issue here is particularly apparent in cultural and affective education. There are three elements to cultural competence: interpretation, responsiveness and appropriate interaction. Emotional intelligence is related to recognising, understanding, and regulating emotional information. Aggregate labels are provided in the supplied data set and the present paper does not make any claims for any particular measurement model. Despite this, there is recent evidence that emotional intelligence is relevant to academic functioning. Several large meta-analyses reported a positive correlation between emotional intelligence and academic performance in various traditions of measuring emotional intelligence (MacCann et al., 2020). A second meta-analysis found a similar relationship in secondary education, and also that effect size is influenced by the instrument and conceptual stream (Sánchez-Álvarez et al., 2020).

Perspective taking and emotional response could be part of these outcomes and may be relevant to virtual and simulated environments, since they can present opportunities for learners to gain exposure to situations where they need to take on a perspective and respond emotionally. A review of virtual-reality empathy interventions showed that virtual experience has a significant effect on both emotional and cognitive aspects of empathy, but that effects vary based on design quality (Ventura et al., 2020). Another applied review in medical and nursing training reveals that simulations can complement experiential learning, but warns against inconsistent evidence quality and inconsistent outcome definitions (Barteit et al., 2021; Plotzky et al., 2021). The results of these assessments reinforce a multi-outcome evaluation; however, they do not warrant the assumption that all domains will benefit to the same extent.

The growing number of discussions around the metaverse and immersive learning has increased the importance of rigorous assessment. Conceptual papers address new opportunities for persistent, social and embodied learning, as well as access, privacy, realism, identity and governance issues (Kye et al., 2021; Mystakidis, 2022; Tlili et al., 2022). More comprehensive evaluations of augmented and virtual reality also point to the fact that adoption is quicker than evidence standards (Al-Ansi et al., 2023). Rather than rejecting technology, its appropriate response is to accept it. It should be used with decision rules that retain uncertainty and expose trade-offs.

2.3 Why separate tests and weighted averages are insufficient

Now suppose that an intervention focuses on two outcomes. A traditional analysis can yield one "significant result" and one "uncertain result. The researcher will then have to make an interpretation. If one test is important, then the entire thing is not effective. Setting the significance level for both outcomes makes the p-value the decision criterion and creates the potential for instability around the p-value. The average of the two effects seems easy, but the average is scale- and weight-dependent. A large effect in one outcome can offset a small or negative effect in another outcome. This compensation may not be consistent with the purpose of the program.

The scale problem is addressed in standardisation, but not the weighting problem. Equal weights remain equal weights, and don't necessarily capture stakeholder priorities. Sometimes explicit weights are justified. In other cases, stakeholders are unable to justify an exact balance amongst domains. In the second case, the minimax criterion comes in handy. It emphasizes "balanced adequacy" over "aggregate reward. It's a conservative criterion, but it's clearly stated. An intervention is as weak as its weakest protected outcome, it says.

The statistical literature also warns against interpreting evidence in isolation as a hit-or-miss test of significance. These factors should be considered when interpreting results, including effect sizes and uncertainty intervals, design assumptions, and data-generating processes (Lakens, 2022; Zhang et al., 2020). Normality tests are not sufficient to select a method and small deviations from normality are often unimportant compared to the presence of heteroskedasticity or dependence (Knief & Forstmeier, 2021). The use of existing data or incompletely documented data also calls for clarity with respect to design and assumptions, particularly when employing causal language (Grosz et al., 2020; Tennant et al., 2021). The concerns inspire the use of the models and their adjustments, the use of robust standard errors, the use of bootstrapped uncertainty, and the use of sensitivity analyses here.

2.4 Minimax and distributionally robust decision-making

A decision is separated from the uncertain quantities that are used to assess the decision in robust optimization. The decision-maker sets the uncertainty set and takes the alternative that works best at the worst case in the set. The extension of this idea is distributionally robust optimization, where there is also uncertainty about the distribution that generates the data (Duchi et al., 2021; Rahimian & Mehrotra, 2022). Ambiguity sets, such as Wasserstein and related ones, have been studied in particular due to their link with empirical data and protected performance of out-of-sample data (Blanchet et al., 2022; Gao & Kleywegt, 2023). More recently, it has been demonstrated that distributional robust rules can be seen as optimal data-driven decisions subject to certain conditions (Van Parys et al., 2021).

The present application is not as complex as a full distributionally robust program, but is simplified on purpose. The design of the data set is too small and too sparse to estimate a complex ambiguity set. The study therefore employs a "confidence bound uncertainty". An intervention contrast and lower confidence limit are provided for each outcome. The strategy that has the lowest minimum lower limit is chosen. A model of the finite scenario type. It maintains the core minimax logic and maintains all the empirical quantities observable and auditable.

In a multistage setting, Shapiro (2021) distinguishes among the three types (risk neutral, risk averse, distributionally robust). The difference is helpful the decisions that are made in the educational context are not exactly those problems. The risk-neutral rule would ensure a maximum average estimated gain. One of the rules might penalize variance and be risk averse. The present rule preserves the bottom limit of each one of the required outcomes, and then optimizes the weakest one. Hence, it is outcome-robust and uncertainty-aware. It doesn't mean that all educational choices must be minimax. It creates a defensible rule when the improvement is a substantive requirement and there is no justification to weigh.

2.5 Research gap and hypotheses

Theoretical gap. The models of learning that are called for as immersive are not usually those that describe how multiple outcomes can be considered together to assess a decision-maker's presence, agency, cognitive processing, and affect. Robust optimization offers a decision rule but this rule has not been very much translated into a clear educational evaluation rule. The gap is located at a pedagogical-mechanism and a multi-outcome decision-theory interface.

Empirical gap. There are inconsistent findings in recent research on VR, and often the emphasis is on achievement, motivation or presence. Cultural and affective endpoints are not as commonly considered as a pair. If more than one outcome is reported, the outcomes are typically interpreted in a parallel fashion, not via a formal balance criterion. The provided data enables a focused demonstration focused on cultural competence and emotional intelligence.

Methodological gap. There are frequent weighted composite scores, and frequently the weights are not theoretically justifiable. Multiple testing and multiple interpretations will arise. There is an alternative that does not involve weights: a lower bound maximin rule. It also delineates the binding outcome and can be carried out in robustness analysis.

Contextual and data-related gap. The dataset includes pre-test and post-test composites, but no pre-test/post-test item information or process information. It also includes PAIRWISE BASELINE EQUALITY. This is not an optimum environment, but methodologically it is instructive. It demonstrates that a minimax evaluation can be distinguished from any unbacked statements of measurement and causal mechanism.

Empirical expectations are expressed as conditions. H1: The pretest-adjusted differences in cultural competence between the EL-VR group and comparison group will be positive. H2: the EL-VR group will have positive pretest-adjusted differences in emotional intelligence. The EL-VR strategy will have a positive worst-outcome standardized lower bound, and thus be chosen by the maximin model. H4: Strategy selection would not change if using alternative estimators, uncertainty procedures and outcome normalizations.

3. Data and Methods

3.1 Study design and data source

The analysis uses the workbook supplied with the study materials. It contains one worksheet and 100 rows of participant-level data. The variables are ID, Group, CC_Pre, EI_Pre, CC_Post, and EI_Post. Group contains two codes, with 50 records in each group. The source materials describe a comparison between traditional instruction and an experiential-learning intervention supported by virtual reality. The numeric mapping is not stored in the workbook. Group 2 is treated as EL-VR and Group 1 as comparison for the present analysis, but the authors must confirm this mapping against the original allocation record.

The observation level is the individual record. The design includes two measurement occasions. Calendar dates, institution, course, instructor, intervention duration, recruitment procedure, and assignment mechanism are not present in the workbook. No demographic or behavioral covariates are available. The analysis therefore focuses on the observed pretest-posttest structure. It does not infer subgroup effects, mediation, moderation, or dose-response relationships.

3.2 Data validation

A complete structural audit was conducted before model estimation. All 100 IDs are unique. Group counts are balanced. No values are missing in the six columns. The four outcome variables are numeric and fall between 2.18 and 5.00. The observed range is compatible with a five-point composite scale, but the instrument documentation is absent. Four cultural-competence posttest values and two emotional-intelligence posttest values equal 5.00, all in Group 2. This small amount of ceiling concentration is considered in the sensitivity analysis.

A critical feature is that the 50 baseline cultural-competence values in Group 1 are exactly repeated, in the same order, in Group 2. The same is true for emotional intelligence. Thus, each row in Group 1 can be paired with a row in Group 2 having an identical two-outcome baseline profile. The pattern may reflect exact matching, copied baseline records, or a data-construction step. The workbook does not identify a pair variable, so a pair index was reconstructed from row order for sensitivity analysis only. The primary analysis does not assume that the two records in a pair are the same person.

The dataset contains composite scores rather than item responses. Internal consistency, factor structure, convergent validity, and discriminant validity cannot be computed. It would be inappropriate to derive Cronbach's alpha from the four composite variables because they represent different constructs and time points. Recent measurement guidance also warns that alpha should not be treated as a universal validity certificate (Hayes & Coutts, 2020). The paper therefore reports score correlations and explicitly limits measurement claims.

3.3 Variables and estimands

Cultural competence at posttest and emotional intelligence at posttest are the two dependent variables. Group membership is the independent variable. The corresponding pretest score is used as a baseline covariate. There are no supported mediators, moderators, or demographic controls. Change scores are calculated for descriptive and sensitivity purposes as posttest minus pretest. Table 1 defines the variables.

Table 1. Variable definitions and analytical roles

Variable	Role	Type	Observed coding range or	Analytical use
ID	Identifier	Integer	1–100	Uniqueness check only
Group	Independent variable	Binary integer	1 or 2	Group 2 coded 1 in models; mapping requires confirmation
CC_Pre	Baseline covariate	Continuous composite	2.37–4.83	Baseline adjustment and standardization

Variable	Role	Type	Observed coding range	Analytical use
EI_Pre	Baseline covariate	Continuous composite	2.18–4.73	Baseline adjustment and standardization
CC_Post	Dependent variable	Continuous composite	2.34–5.00	Primary cultural-competence outcome
EI_Post	Dependent variable	Continuous composite	2.21–5.00	Primary emotional-intelligence outcome
CC_Change	Derived outcome	Continuous	-0.13–1.16	Robustness and descriptive analysis
EI_Change	Derived outcome	Continuous	-0.21–1.14	Robustness and descriptive analysis

Note. No mediator, moderator, demographic control, or process variable is present in the workbook.

The primary estimand for outcome k is the pretest-adjusted mean difference between Group 2 and Group 1. For participant i , strategy j , and outcome k , the observed change is defined as:

$$\Delta_{ijk} = Y_{ijk,post} - Y_{ijk,pre}. \quad (1)$$

The adjusted model for each outcome is:

$$Y_{ik,post} = \alpha_k + \tau_k T_i + \beta_k Y_{ik,pre} + \epsilon_{ik}, \quad (2)$$

where T_i equals one for Group 2 and zero for Group 1. The coefficient τ_k is the adjusted group contrast. Heteroskedasticity-consistent HC3 standard errors are used because change-score variances differ strongly between the groups. A multivariate analysis of covariance is also reported as a joint descriptive test across the two posttest outcomes.

3.4 Standardization and uncertainty

The minimax model requires comparable outcome units. Each adjusted contrast is divided by the baseline standard deviation for the corresponding outcome:

$$g_{\hat{jk}} = \tau_{\hat{jk}} / s_{k,pre}. \quad (3)$$

This standardization is interpretable as the adjusted gain in baseline standard-deviation units. It also makes the decision invariant to a positive change in the original measurement units. The comparison strategy is treated as the reference with incremental gain zero. The EL-VR strategy has the two estimated standardized contrasts.

Uncertainty is represented by a lower confidence bound. Let L_{jk} denote the lower bound for strategy j and outcome k . For the empirical model, the bounds are estimated with a pair-based nonparametric bootstrap that resamples the 50 observed baseline pairs and recomputes the difference in mean change. Twenty thousand resamples are generated with a fixed seed. The lower 2.5th percentile is used as a conservative bound. HC3-based intervals are reported in parallel.

3.5 The minimax optimization model

Consider J instructional strategies and K required educational outcomes. Let x_j indicate whether strategy j is selected. Let L_{jk} be the protected standardized gain for strategy j on outcome k . The pure-strategy maximin problem is:

$$\text{maximize}_{(x,t)} \quad t \quad (4a)$$

$$\text{subject to} \quad t \leq \sum_{(j=1)^J} L_{jk} x_j, \quad \text{for every } k = 1, \dots, K, \quad (4b)$$

$$\sum_{(j=1)^J} x_j = 1, \quad x_j \in \{0,1\}. \quad (4c)$$

The scalar t is the protected performance of the weakest outcome. In the empirical application, J equals two and K equals two. The comparison strategy has $L_{0k} = 0$ because the objective is incremental gain over the comparison. The EL-VR strategy has bootstrap lower bounds for cultural competence and emotional intelligence.

Proposition 1. For a finite set of pure strategies, the solution of (4) is any strategy j^* that maximizes $\min_k L_{jk}$.

Proof. If x_j is binary and sums to one, exactly one strategy is selected. For a selected strategy j , constraints (4b) imply t cannot exceed its smallest protected outcome, $\min_k L_{jk}$. Setting t equal to that minimum is feasible. Choosing the strategy with the largest such minimum therefore solves the program.

Proposition 2. If each outcome is transformed by a positive affine change of units and the effect is standardized by the corresponding baseline standard deviation, the maximin strategy is unchanged.

Proof. A positive affine transformation multiplies the adjusted contrast and the baseline standard deviation by the same positive constant. Their ratio is unchanged. The protected standardized vector and the objective value are therefore unchanged.

A mixed-strategy relaxation could allow x_j between zero and one. That formulation would represent a resource portfolio across instructional modes. The current dataset does not contain treatment dose or mixed-delivery outcomes, so linear interpolation would be unsupported. Only pure strategies are evaluated.

3.6 Minimax regret sensitivity

A second decision rule measures regret relative to the best protected strategy for each outcome. Let $B_k = \max_j L_{jk}$. The minimax regret model is:

minimize (x, R) R , subject to $R \geq B_k - \sum_j L_{jk} x_j$ for every k . (5)

This criterion asks which strategy has the smallest largest shortfall from the outcome-specific best strategy. If one strategy protectedly dominates the other across all outcomes, the maximin and minimax-regret rules agree. The empirical analysis checks this condition.

3.7 Statistical and sensitivity procedures

The primary univariate analyses use HC3 analysis of covariance. The joint outcome analysis uses Wilks' lambda in a multivariate model with both baseline variables. Robustness procedures include Welch tests on change scores, paired t tests based on the reconstructed baseline pairs, Wilcoxon signed-rank tests, Mann-Whitney tests, 20% trimmed-mean comparisons, pair-clustered ANCOVA, nonparametric bootstrap intervals, and leave-one-pair-out calculations. A headroom normalization divides the adjusted contrast by the distance between the baseline mean and the observed scale maximum of five. This checks whether the decision depends on standard-deviation scaling.

Machine learning is not used. The dataset has only 100 observations, one binary group indicator, and four composite scores. A predictive model with cross-validation and hyperparameter tuning would not answer the decision question and would create avoidable overfitting risk. The analytical design therefore favors transparent estimands and exact decision rules.

All calculations were performed in Python 3.13.5 with NumPy 2.3.5, SciPy 1.17.0, statsmodels 0.14.6, and Matplotlib 3.10.8. Effect sizes and uncertainty intervals are interpreted together rather than through p values alone, consistent with current reporting guidance (Ben-Shachar et al., 2020; Lakens, 2022; Lüdtke et al., 2021; Zhang et al., 2020). Figure 2 presents the full workflow.

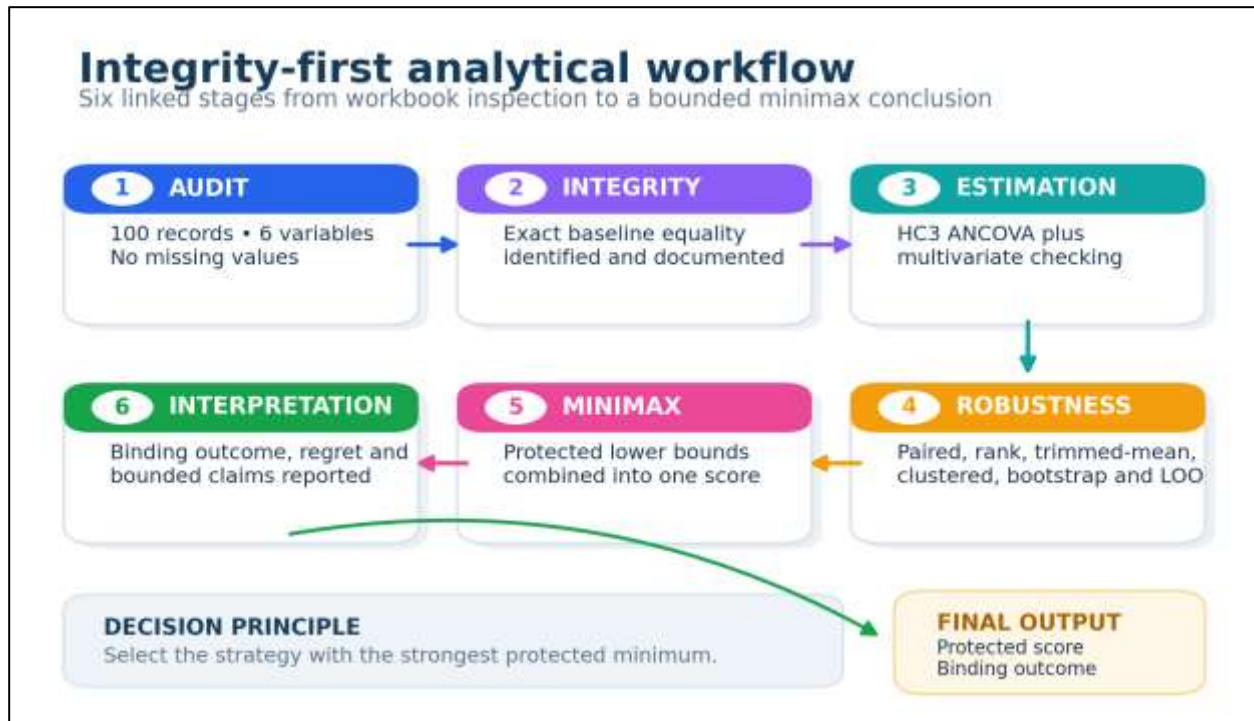


Figure 2. Integrity-first data validation and analytical workflow.

4. Results

4.1 Data completeness and baseline structure

The dataset contains 100 complete records. Group 1 and Group 2 each contain 50 observations. There are no missing values, duplicate IDs, or nonnumeric outcome entries. Cultural competence ranges from 2.37 to 4.83 at pretest and from 2.34 to 5.00 at posttest. Emotional intelligence ranges from 2.18 to 4.73 at pretest and from 2.21 to 5.00 at posttest.

The baseline means are identical across groups because the entire baseline vectors are pairwise repeated. Cultural competence has a mean of 3.603 and a standard deviation of 0.474 in each group. Emotional intelligence has a mean of 3.676 and a standard deviation of 0.444 in each group. This is stronger than ordinary baseline equivalence. It is exact profile equality. The feature removes observed baseline mean imbalance but raises a provenance question that cannot be resolved from the workbook alone.

At posttest, Group 1 has a mean cultural-competence score of 3.623 (SD = 0.497), while Group 2 has a mean of 4.226 (SD = 0.489). The corresponding emotional-intelligence means are 3.687 (SD = 0.463) and 4.286 (SD = 0.454). Mean change in Group 1 is 0.020 for cultural competence and 0.011 for emotional intelligence. Mean change in Group 2 is 0.622 and 0.610, respectively. Table 2 reports the descriptive statistics, and Figure 3 compares the pretest-to-posttest mean profiles across groups.

Table 2. Descriptive statistics by outcome and group

Outcome	Group	Pretest M (SD)	Posttest M (SD)	Change M (SD)
Cultural competence	Comparison	3.603 (0.474)	3.623 (0.497)	0.020 (0.083)
Cultural competence	EL-VR	3.603 (0.474)	4.226 (0.489)	0.622 (0.273)
Emotional intelligence	Comparison	3.676 (0.444)	3.687 (0.463)	0.011 (0.109)
Emotional intelligence	EL-VR	3.676 (0.444)	4.286 (0.454)	0.610 (0.237)

Note. Each group contains 50 records. Change equals posttest minus pretest.

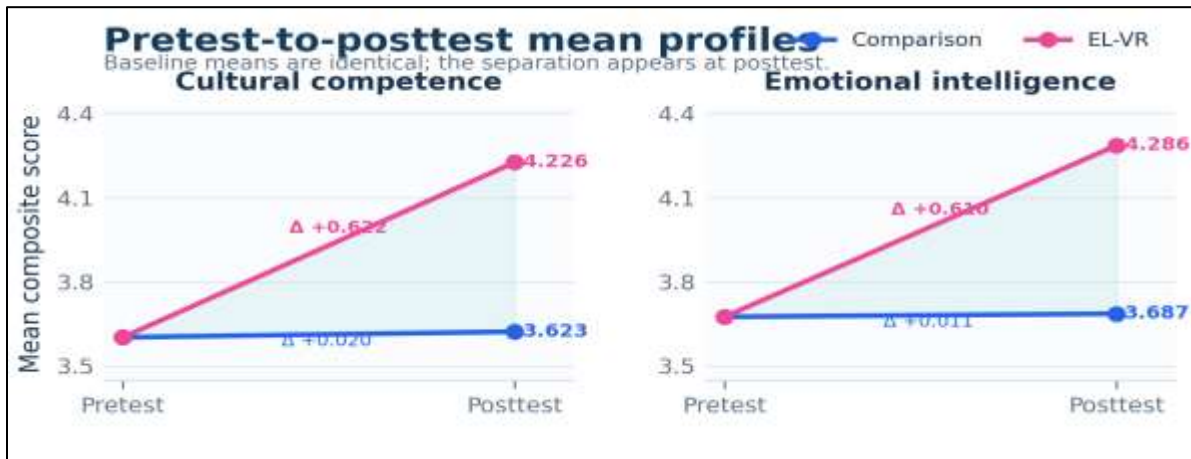


Figure 3. Pretest-to-posttest mean profiles by outcome and instructional group.

The overall correlations show strong continuity within constructs. Cultural competence pretest correlates 0.778 with cultural competence posttest, and emotional intelligence pretest correlates 0.765 with emotional intelligence posttest. Cross-construct baseline correlation is -0.030. The posttest correlation is 0.208, partly reflecting common group separation. Within the comparison group, pretest-posttest correlations are 0.987 for cultural competence and 0.972 for emotional intelligence. Within Group 2, they are 0.840 and 0.861. These values indicate score stability, not internal consistency or construct validity. Table 3 and Figure 4 summarize the correlation structure.

Table 3. Pearson correlations among supplied composite scores

Variable	1	2	3	4
1. CC pre	1.000			
2. EI pre	-0.030	1.000		
3. CC post	0.778	-0.064	1.000	
4. EI post	-0.062	0.765	0.208	1.000

Note. $N = 100$. These correlations do not establish internal consistency or construct validity. Item-level data were not supplied.

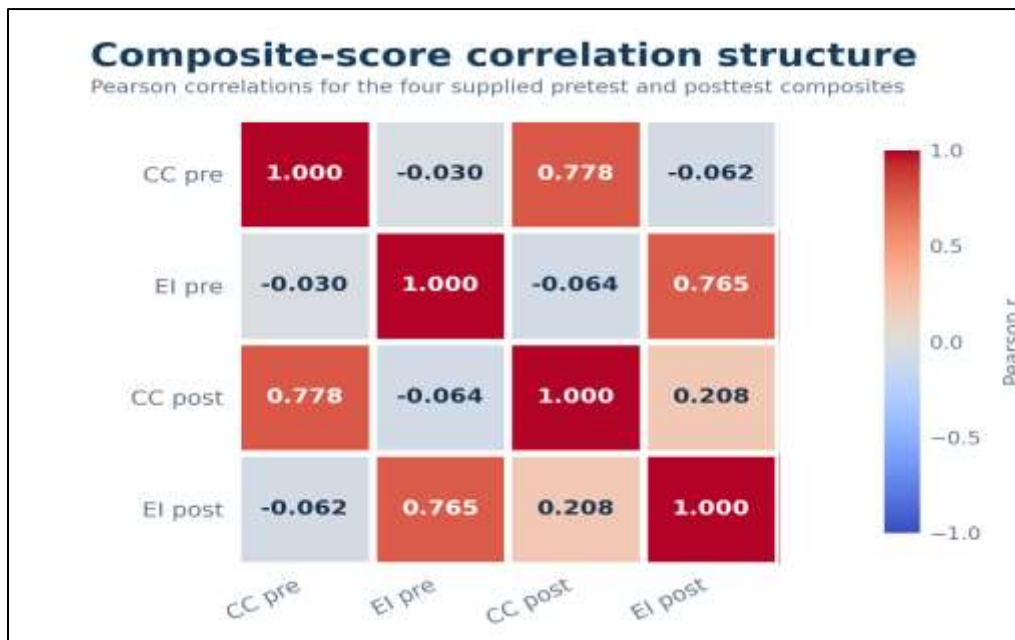


Figure 4. Correlation structure of the supplied composite scores.

4.2 Pretest-adjusted outcome contrasts

The multivariate model shows a strong adjusted group difference across the two posttest outcomes, Wilks' lambda = 0.161, $F(2, 95) = 248.50$, $p < .001$. The result is descriptive of joint separation. It does not resolve the assignment and provenance limitations.

For cultural competence, the HC3-adjusted Group 2 coefficient is 0.602 (SE = 0.041, 95% CI [0.522, 0.683], $p < .001$). The model R squared is 0.881. The classical partial eta squared for the group term is 0.698. For emotional intelligence, the adjusted coefficient is 0.600 (SE = 0.037, 95% CI [0.527, 0.673], $p < .001$). The model R squared is 0.889, and the classical partial eta squared is 0.733. Table 4 reports the models, and Figure 5 displays the adjusted contrasts.

Table 4. Pretest-adjusted outcome models

Outcome	Adjusted difference	HC3 SE	95% CI	z	p	R ²	Partial η^2
Cultural competence	0.602	0.041	[0.522, 0.683]	14.72	< .001	0.881	0.698
Emotional intelligence	0.600	0.037	[0.527, 0.673]	16.07	< .001	0.889	0.733

Note. Adjusted difference is Group 2 minus Group 1. HC3 standard errors are heteroskedasticity-consistent. Partial eta squared is based on the corresponding classical ANCOVA decomposition.

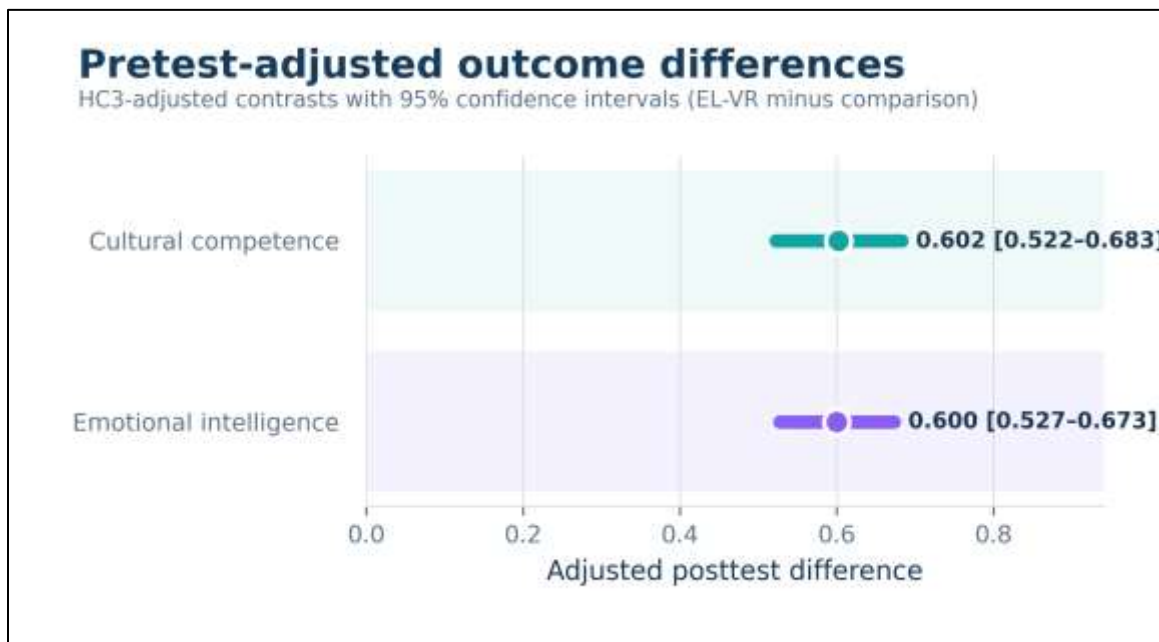


Figure 5. Pretest-adjusted outcome differences with 95% confidence intervals.

The corresponding change-score contrasts are 0.602 for cultural competence and 0.600 for emotional intelligence. Welch tests are significant for both outcomes, with $t(57.93) = 14.95$ and $t(68.66) = 16.25$. The unequal change variances are substantial, which supports the use of heteroskedasticity-robust procedures. The observed group difference therefore does not depend on an equal-variance assumption.

4.3 Minimax solution

The adjusted cultural-competence contrast divided by its baseline standard deviation is 1.271. The adjusted emotional-intelligence contrast divided by its baseline standard deviation is 1.350. Thus, the point-estimate maximin score is $\min(1.271, 1.350) = 1.271$. Cultural competence is the binding outcome. The pair-based bootstrap produces a 95% interval of [1.087, 1.447] for the standardized cultural-competence contrast and [1.187, 1.506] for emotional intelligence when the original baseline standard

deviations are used as calibration constants. The protected vector is therefore (1.087, 1.187). The maximin lower-bound value is 1.087. The comparison reference has an incremental protected vector of (0, 0). Model (4) selects the EL-VR strategy because 1.087 exceeds zero.

The same strategy minimizes maximum regret. The EL-VR lower bound is higher than the comparison reference for both outcomes. Its outcome-specific regret is therefore zero, while the comparison strategy has positive regret in both domains. Figure 6 shows the decision geometry, and Table 5 reports the optimization inputs and solution.

Table 5. Minimax optimization inputs and protected solution

Quantity	Cultural competence	Emotional intelligence
Adjusted raw difference	0.6024	0.5998
Baseline SD	0.4739	0.4442
Standardized point gain	1.2711	1.3502
Pair-bootstrap 95% interval	[1.0871, 1.4466]	[1.1872, 1.5064]
Protected lower bound	1.0871	1.1872
Headroom-normalized point gain	43.13%	45.30%
Headroom lower bound	36.75%	39.42%

Note. The point maximin value is 1.2711. The protected maximin value is 1.0871. Cultural competence is the binding outcome. The comparison reference vector is (0, 0).

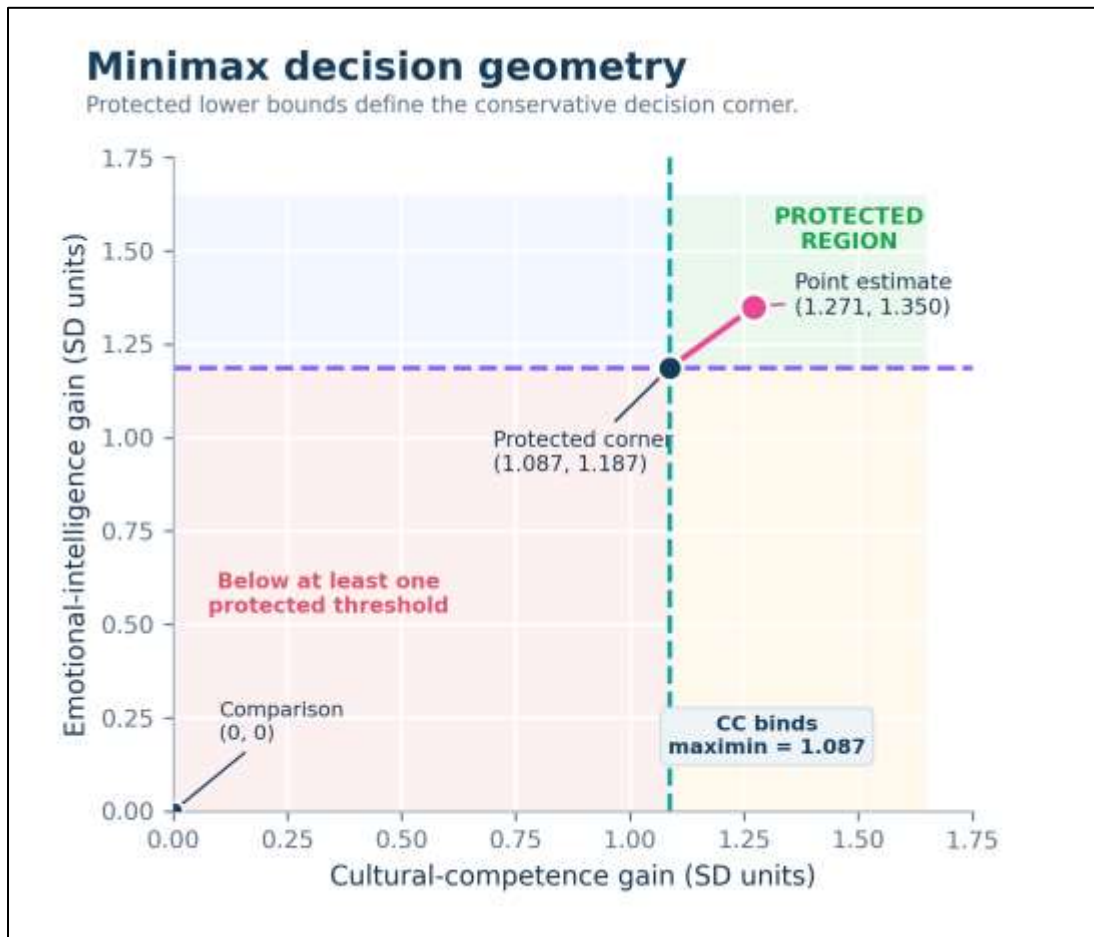


Figure 6. Minimax decision geometry under outcome uncertainty.

4.4 Robustness and sensitivity

The strategy choice remains unchanged under every planned robustness procedure. Pair-clustered ANCOVA produces a cultural-competence interval of [0.515, 0.690] and an emotional-intelligence interval of [0.528, 0.672]. Paired t tests based on the baseline pair index give $t(49) = 13.60$ for cultural

competence and $t(49) = 16.44$ for emotional intelligence. Wilcoxon signed-rank tests are also significant for both outcomes. Independent rank tests lead to the same conclusion.

Twenty-per cent trimmed means yield group contrasts of 0.615 for cultural competence and 0.607 for emotional intelligence. Both are significant. Across 50 leave-one-pair-out analyses, the cultural-competence contrast ranges from 0.591 to 0.617, and the emotional-intelligence contrast ranges from 0.591 to 0.614. The standardized maximin point value ranges from 1.248 to 1.303. No single baseline pair determines the selection.

Headroom normalization produces adjusted gains equal to 43.1% of the available cultural-competence range and 45.3% of the emotional-intelligence range. Pair-bootstrap lower bounds are 36.8% and 39.4%. The protected minimum remains positive and cultural competence remains binding. Table 6 summarizes the robustness results. H1 through H4 are therefore supported within the limits of the supplied data.

Table 6. Robustness and sensitivity analysis

Procedure	Cultural competence	Emotional intelligence	Decision
Welch change-score test	Diff. 0.602; $t = 14.95$	Diff. 0.600; $t = 16.25$	EL-VR
Pair-based t test	$t(49) = 13.60$	$t(49) = 16.44$	EL-VR
Wilcoxon signed-rank	$p < .001$	$p < .001$	EL-VR
20% trimmed means	Diff. 0.615	Diff. 0.607	EL-VR
Pair-clustered ANCOVA CI	[0.515, 0.690]	[0.528, 0.672]	EL-VR
Leave-one-pair-out range	0.591–0.617	0.591–0.614	EL-VR
Headroom lower bound	36.75%	39.42%	EL-VR

Note. All p values for the listed parametric and rank-based tests are below .001 unless an interval is shown.

5. Discussion

5.1 Interpretation of the findings

If the empirical demonstration leads to a clear mathematical decision, then that is the one selected. Group 2 shows positive adjusted contrasts for both outcomes given. The lower bound, after the standardization and uncertainty protection, remains positive. So the EL-VR strategy is selected by the maximin rule. The result is a protected standardized gain, specifically for cultural competence, which is less than the emotional-intelligence gain, thus making it the binding outcome.

The binding result is not a fail. It is the result which constrains the program's guaranteed performance. The distinction is helpful for instructional design. A weighted average would give a standardized gain of approximately 1.31 and would not be an indicator of a domain that lacks balanced performance. In the maximin model, cultural competence is directly identified. Instructional resources can be focused on the tasks, prompts, or feedback processes that are most salient to cultural interpretation and transfer, if the intervention is developed further.

The findings are in line with recent studies conducted in the virtual reality space that found immersive environments could help learning when integrated into a structured pedagogy (Hamilton et al., 2021; Makransky & Petersen, 2021; Radianti et al., 2020). They may also be applied to evidence showing that experiential designs can affect affective processes and perspective taking, which are only meaningful in the context of a good learning task (Barteit et al., 2021; Parong & Mayer, 2021; Ventura et al., 2020). Presence, cognitive load, reflection, collaboration and empathy are not measured in the present data. However, the analysis is not able to determine which process caused the separation seen. Only tests the provided outcome pattern.

5.2 Theoretical contribution

The primary theoretical contribution is the link between the theory of multi-outcome learning and the minimax decision-making. Theories about mechanisms are called the theory of immersive learning. They do not usually include a formal rule to help decide if an integrated intervention is sufficient for multiple outcomes. The maximin model incorporates that one. It takes each of the outcomes that it has been designed to meet as a scenario that the chosen strategy needs to withstand.

The model also provides a better understanding of the importance of standardization. Minimax decisions may be misled in case of different units of outcomes. When dividing by baseline standard deviation, the dependence on the units is eliminated and the proposition 2 is upheld. This does not assert that in all domains one s.d. represents the same social value. It only sets up a common statistical scale. If the stakeholders are able to assign defensible preference weights, a weighted multiobjective model might be more appropriate. If they do not, there is a sensible default: the maximin rule.

Another theoretical work is the use of protected instead of expected outcome vectors. A key feature of robust optimization is that the uncertainty in the estimated performance should be considered for making a decision (Duchi et al., 2021; Rahimian & Mehrotra, 2022; Van Parys et al., 2021). The uncertainty set constructed in the present model is made up of bootstrap confidence limits. This is easier than the Wasserstein ambiguity set, but follows the same concepts. The decision is not solely based upon the best point estimate, but rather on what can be protected.

A finite-scenario structure is also apparent. Some educational research studies only have a limited number of strategies and outcomes, which limits the ability to use a complex optimization model. In that context, model (4) is still of use. Two strategies can be solved by inspection, many strategies are solved by mixed-integer programming and when there is a justified portfolio interpretation, a strategy can be solved by linear programming. Despite the empirical illustration being small, the mathematical form is general.

5.3 Methodological contribution

The analysis proves to have an integrity-first workflow. The initial profiles are identical between groups. A standard report would merely say that the groups were matched prior to the treatment. That's not quite true. It is very rare that the equality holds on 100 base points and needs to be explained. The study avoids ordinary random balance by documenting the pattern and testing pair-based dependence – if the data indicate otherwise.

Measurement, estimation and decision are also differentiated in the primary and sensitivity models. The issue of validity of CC and EI as composites is a measurement issue. Without the information at the item-level, that can't be answered. Estimation is related to the modified group differences and the uncertainty of those differences. Those quantities can be computed. Decision involves which one of the two strategies yields the greatest protected minimum. The minimax model does not assume that the missing measurement evidence exists, it is just estimated.

This separation proves to be helpful to limit a common mistake in the research of TE learning. The claims of attractiveness of the intervention, a statistically significant coefficient, and a valid construct are different. Different evidence is needed. The present results lend support to a high observed group contrast in the provided composites. They do not provide a psychometric evaluation of the scales, a separate effect of the VR or long-term retention.

The methodological contribution involves not using machine learning. Sometimes, a decision needs to be made, or an appropriate sample size is not met in educational studies and predictive algorithms are incorporated. The core question here is low-dimensional and is causal-adjacent. It is directly answered by transparent ANCOVA, uncertainty of the bootstrap, and a small robust program. Random forests or neural networks would decrease interpretability, and would not produce variables that are not present in the data.

5.4 Practical implications

The model helps the educator make a simple decision aid. First, set the goals that should be accomplished by an intervention. Second, make a best guess of each possible outcome on a similar scale. Thirdly, anchor the estimates with believable lower limits. Finally, choose the strategy offering the maximum minimum protected value. More information is provided than a list of p values since the educational bottleneck is made explicit.

Designed for virtual-reality learning, the binding result can be used to inform the iterative design process. Within this data set, cultural competence is bonded. The design cycle could be reinforced in a subsequent design cycle with opportunities for contextual interpretation, guided comparison, reflective

prompts, or cultural understanding. This objective would not be to raise a global average. It would be a bad idea to increase the protected minimum while decreasing the emotional-intelligence performance. The approach is helpful for institutional decision makers in scaling with caution. A positive maximin lower bound is indicative of the fact that the observed intervention contrast is not due to one high endpoint and one low endpoint. It fails to address equity, instructor training, hardware reliability, or cost questions, or accessibility questions. If appropriate data is available these criteria can be included as further outcomes. This would make the decision more conservative, but more in line with actual implementation.

The model supports preregistered outcome sets for researchers. It is only useful to perform a minimax analysis if the required domains are specified prior to the results. If outcomes are added during analysis or removed from the analysis, the binding scenario will change. In future applications, the outcome set, the concept of normalization, the level of uncertainty, and the pure/mixed decision space should be defined before.

5.5 Comparison with prior evidence

Results of the observed learning effect corroborate the results of previous meta-analyses that found that immersive and simulated environments are effective at enhancing learning and training outcomes (Kaplan et al., 2021; Villena-Taranilla et al., 2022; Wu et al., 2020). It is greater than many 'pooled effects' that are summarized in general reviews. This disparity cannot be taken as an indicator of high quality of intervention. Available data at the present time is without independent replication, with the same baseline matched, with a small sample, and unclear assignment. These features can give a cleaner contrast than is seen in normal practice.

The emotional-intelligence score is plausible given the evidence that one's experiences, which are emotionally and socially situated, can have some affect on the emotional outcome. However, this shouldn't be taken as solely because they're immersing themselves. There are differences in perspective-taking and empathy effects from design to design (Ventura et al., 2020). The provided information does not distinguish between virtual exposure and collaboration, guidance, novelty, instructor influence or reflective activity. The correct interpretation is that higher posttest composite scores are observed with the EL-VR package.

The cultural-competence outcome is also constrained. Immersive heritage or cultural environments can provide context and embodied exploration, but visual realism doesn't necessarily equate to cultural understanding. Interpretation, alternative viewpoints and teacher scaffolding are still relevant. The advantage of the maximin result is that it does not make the assumption that the two domains react in exactly the same way. It acknowledges that cultural competence is the weaker protected domain, although it estimates that it has a significant benefit.

5.6 Policy and governance implications

Immersive education should be considered as a portfolio of outcomes at policy level, not a technology purchase. Decisions on procurement are sometimes made based on device capabilities, engagement or demonstration value. A strong outcome framework focuses on delivering protected educational benefit. It may also include access, cost, privacy, cybersickness, and instructor workload as other scenarios when they can be added in.

Issues of identity, data, inclusion, and platform control are identified in the metaverse literature (Kye et al., 2021; Tlili et al., 2022). These discussions can be aided by a minimax approach, which will not easily be persuaded to compensate. For example, the learning gain should not be sufficient to justify an inadequate accessibility result. The decision-maker can specify accessibility as one of the scenarios, and set a positive protected threshold.

The present empirical model does not contain any policy variables; thus there is no direct policy effect estimated. The implication is methodological. Requirements may be non compensatory and institutions may assess candidate strategies by reference to these requirements. This is better to do if some outcomes are protections and not wishes.

6. Limitations and Future Research

There are limitations to the study. One, labels with numbers are not recorded in the workbook. The interpretation of Group 2 is based on the following narrative and pattern, but must be verified from the original allocation file. The intervention protocol and group coding rule should be included in a publication version.

Secondly, the baseline profiles are exactly the same between groups. If this is a planned exact matching, then the variables matched, the exact matching algorithm, the timing of the exact matching, and the identifiers of the pairs being matched should be reported. If it shows evidence of data copying, or the construction of scores, then this reason should be explained and the raw data should be checked. Until provenances can be identified, causal interpretation should be done with caution.

Third, there is no detail of the assignment. The analysis does not provide any information on whether or not the groups were randomly formed, formed from intact classes, or self-selected or formed by other means. Only the observed baseline composite is adjusted for in a pretest. Does not exclude instructor or class differences, motivation or exposure differences. The results should thus be interpreted as adjusted group contrasts and not as treatment effects.

4) Only aggregate composite scores are provided. The reliability of items, factor validity, and response pattern diagnostics are not available. Outcome labels are taken as given. The full instrument, scoring rules, subdimensions and evidence for the instrument's validation should be provided in future research. It should also assess the model stability of the measurement model across groups and time.

Fifth, the components of the intervention are not measured. The model assesses strategies, not immersion, guided tasks, collaboration or reflection alone. Instructional architecture optimization should be carried out in a factorial or component study. Similarly, no mediation, moderation or analysis of feature importance can be conducted based on the current data.

Sixth, it is based on an unspecified institution and time period, and the sample is small. No delayed post-test. The analysis does not allow for the conclusions about durability, transfer and external validity. Effective replication in multiple sites with larger samples and preregistered outcomes is required.

Seventh is that uncertainty set is sampling intervals-based. It lacks modelling uncertainty, measurement error, institution-to-institution variation in implementation and distribution shift among institutions. In the future, hierarchical robust optimization, Wasserstein ambiguity sets or Bayesian posterior lower bounds may be helpful when more detailed data are available (Blanchet et al., 2022; Gao & Kleywegt, 2023).

Eighth, the model is based on a non-compensatory objective. This is appropriate in the presence of balanced outcome adequacy. It could be too conservative if stakeholders can justify specific trade-offs. Later research should compare the maximin, minimax regret, weighted utility, epsilon-constraint and Pareto-front approaches in the context of prespecified decision preferences.

Implementation criteria should be added to future research as well. Additional outcomes or constraints can be cost, instructor time, access to hardware, cybersickness, and data governance. This would shift the focus from an efficacy comparison to a comprehensive institutional decision model.

7. Conclusion

This study proposes a minimax optimization model that is used to assess virtual reality experiential learning from various educational outcomes. The model standardizes the contrast of the observed outcome, guarantees them lower uncertainty bounds and chooses the strategy that maximizes the lowest bound. As such it doesn't have arbitrary outcome weights, and it doesn't have a good result in one area that masks a bad result in another.

The adjusted posttest advantages for cultural competence and emotional intelligence are 0.602 and 0.600, respectively, for Group 2 in the provided data set. The standardized gains are 1.271 and 1.350. The vectors protected by the conservatives are 1.087 and 1.187 and the maximin is 1.087. The result is cultural competence. Change score, pair-based, rank-based, trimmed mean, clustered, bootstrap, leave one pair out and headroom normalized all result in the same strategy.

The mathematical end result is clear, but the empirical claim needs to be “tethered.” There is no information in the workbook about the exact matching of the items to baseline, the assignment information, the intervention information, or the calendar context. It should thus be understood that the article is not full causal evidence of a virtual-reality program, but is a “compliant” use of a general decision model.

The value of the framework is the decision logic, which is broader than the framework itself. If an educational intervention is required to have an impact on multiple domains and is not required to have an impact on any domain that is protected, then a meaningful criterion is a meaningful criterion. This criterion is made explicit, reproducible, and open to sensitivity analysis in a minimax model.

Ethics Statement

Ethics approval and informed-consent information were not included in the supplied materials. The authors must insert the approved protocol identifier, approving body, consent procedure, and any applicable waiver before submission.

Data Availability Statement

The analyzed workbook was supplied by the study authors and is not publicly archived in the materials available for this analysis. The authors should state the access conditions and provide a repository link or justified restriction before submission.

Conflict of Interest

No conflict-of-interest information was supplied. Each author must complete this statement before submission.

Funding

No funding information was supplied. The authors must identify all funding sources and grant numbers, or state that the work received no specific funding.

Author Contributions

Author names and contribution roles were not supplied. A CRediT contribution statement should be completed on the separate title page.

REFERENCES

1. Al-Ansi, A. M., Jaboob, M., Garad, A., & Al-Ansi, A. (2023). Analyzing augmented reality (AR) and virtual reality (VR) recent development in education. *Social Sciences & Humanities Open*, 8, 100532. <https://doi.org/10.1016/j.ssaho.2023.100532>
2. Barteit, S., Lanfermann, L., Bärnighausen, T., Neuhaus, F., & Beiersmann, C. (2021). Augmented, mixed, and virtual reality-based head-mounted devices for medical education: Systematic review. *JMIR Serious Games*, 9(3), e29080. <https://doi.org/10.2196/29080>
3. Ben-Shachar, M. S., Lüdtke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, 5(56), 2815. <https://doi.org/10.21105/joss.02815>
4. Blanchet, J., Murthy, K., & Si, N. (2022). Confidence regions in Wasserstein distributionally robust estimation. *Biometrika*, 109(2), 295–315. <https://doi.org/10.1093/biomet/asab026>
5. Bond, M., Bedenlier, S., Marín, V. I., & Händel, M. (2021). Emergency remote teaching in higher education: Mapping the first global online semester. *International Journal of Educational Technology in Higher Education*, 18, 50. <https://doi.org/10.1186/s41239-021-00282-x>
6. Checa, D., & Bustillo, A. (2020). A review of immersive virtual reality serious games to enhance learning and training. *Multimedia Tools and Applications*, 79, 5501–5527. <https://doi.org/10.1007/s11042-019-08348-9>
7. Di Natale, A. F., Repetto, C., Riva, G., & Villani, D. (2020). Immersive virtual reality in K-12 and higher education: A 10-year systematic review of empirical research. *British Journal of Educational Technology*, 51(6), 2006–2033. <https://doi.org/10.1111/bjet.13030>

8. Duchi, J. C., Glynn, P. W., & Namkoong, H. (2021). Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3), 946–969. <https://doi.org/10.1287/moor.2020.1085>
9. Gao, R., & Kleywegt, A. J. (2023). Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2), 603–655. <https://doi.org/10.1287/moor.2022.1275>
10. Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, 15(5), 1243–1255. <https://doi.org/10.1177/1745691620921521>
11. Hamilton, D., McKechnie, J., Edgerton, E., & Wilson, C. (2021). Immersive virtual reality as a pedagogical tool in education: A systematic literature review of quantitative learning outcomes and experimental design. *Journal of Computers in Education*, 8, 1–32. <https://doi.org/10.1007/s40692-020-00169-2>
12. Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But... *Communication Methods and Measures*, 14(1), 1–24. <https://doi.org/10.1080/19312458.2020.1718629>
13. Kaplan, A. D., Cruik, J., Endsley, M., Beers, S. M., Sawyer, B. D., & Hancock, P. A. (2021). The effects of virtual reality, augmented reality, and mixed reality as training enhancement methods: A meta-analysis. *Human Factors*, 63(4), 706–726. <https://doi.org/10.1177/0018720820904229>
14. Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, 53, 2576–2590. <https://doi.org/10.3758/s13428-021-01587-5>
15. Kye, B., Han, N., Kim, E., Park, Y., & Jo, S. (2021). Educational applications of metaverse: Possibilities and limitations. *Journal of Educational Evaluation for Health Professions*, 18, 32. <https://doi.org/10.3352/jeehp.2021.18.32>
16. Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, 8(1), 33267. <https://doi.org/10.1525/collabra.33267>
17. Lüdtke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R package for assessment, comparison and testing of statistical models. *Journal of Open Source Software*, 6(60), 3139. <https://doi.org/10.21105/joss.03139>
18. MacCann, C., Jiang, Y., Brown, L. E. R., Double, K. S., Bucich, M., & Minbashian, A. (2020). Emotional intelligence predicts academic performance: A meta-analysis. *Psychological Bulletin*, 146(2), 150–186. <https://doi.org/10.1037/bul0000219>
19. Makransky, G., & Petersen, G. B. (2021). The Cognitive Affective Model of Immersive Learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality. *Educational Psychology Review*, 33, 937–958. <https://doi.org/10.1007/s10648-020-09586-2>
20. Morris, T. H. (2020). Experiential learning: A systematic review and revision of Kolb's model. *Interactive Learning Environments*, 28(8), 1064–1077. <https://doi.org/10.1080/10494820.2019.1570279>
21. Mystakidis, S. (2022). Metaverse. *Encyclopedia*, 2(1), 486–497. <https://doi.org/10.3390/encyclopedia2010031>
22. Parong, J., & Mayer, R. E. (2021). Cognitive and affective processes for learning science in immersive virtual reality. *Journal of Computer Assisted Learning*, 37(1), 226–241. <https://doi.org/10.1111/jcal.12482>
23. Petersen, G. B., Klingenberg, S., Mayer, R. E., & Makransky, G. (2020). The virtual field trip: Investigating how to optimize immersive virtual learning in climate change education. *British Journal of Educational Technology*, 51(6), 2099–2115. <https://doi.org/10.1111/bjet.12991>
24. Plotzky, C., Lindwedel, U., Sorber, M., Loessl, B., König, P., Kunze, C., Kugler, C., & Meng, M. (2021). Virtual reality simulations in nurse education: A systematic mapping review. *Nurse Education Today*, 101, 104868. <https://doi.org/10.1016/j.nedt.2021.104868>

25. Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education*, 147, 103778. <https://doi.org/10.1016/j.compedu.2019.103778>
26. Rahimian, H., & Mehrotra, S. (2022). Distributionally robust optimization: A review. *Open Journal of Mathematical Optimization*, 3, 4. <https://doi.org/10.5802/ojmo.15>
27. Rapanta, C., Botturi, L., Goodyear, P., Guàrdia, L., & Koole, M. (2020). Online university teaching during and after the COVID-19 crisis: Refocusing teacher presence and learning activity. *Postdigital Science and Education*, 2, 923–945. <https://doi.org/10.1007/s42438-020-00155-y>
28. Sánchez-Álvarez, N., Berrios Martos, M. P., & Extremera, N. (2020). A meta-analysis of the relationship between emotional intelligence and academic performance in secondary education: A multi-stream comparison. *Frontiers in Psychology*, 11, 1517. <https://doi.org/10.3389/fpsyg.2020.01517>
29. Shapiro, A. (2021). Tutorial on risk neutral, distributionally robust and risk averse multistage stochastic programming. *European Journal of Operational Research*, 288(1), 1–13. <https://doi.org/10.1016/j.ejor.2020.05.047>
30. Tennant, P. W. G., Murray, E. J., Arnold, K. F., Berrie, L., Fox, M. P., Gadd, S. C., Harrison, W. J., Keeble, C., Ranker, L. R., Textor, J., Tomova, G. D., Gilthorpe, M. S., & Ellison, G. T. H. (2021). Use of directed acyclic graphs to guide analyses of causal questions: A review. *International Journal of Epidemiology*, 50(2), 620–632. <https://doi.org/10.1093/ije/dyaa213>
31. Tlili, A., Huang, R., Shehata, B., Liu, D., Zhao, J., Metwally, A. H. S., Wang, H., Denden, M., Bozkurt, A., Lee, L. H., Beyoglu, D., Altinay, F., Sharma, R. C., Altinay, Z., Li, Z., Liu, J., Ahmad, F., Hu, Y., Salha, S., & Burgos, D. (2022). Is Metaverse in education a blessing or a curse: A combined content and bibliometric analysis. *Smart Learning Environments*, 9, 24. <https://doi.org/10.1186/s40561-022-00205-x>
32. Van Parys, B. P. G., Mohajerin Esfahani, P., & Kuhn, D. (2021). From data to decisions: Distributionally robust optimization is optimal. *Management Science*, 67(6), 3387–3402. <https://doi.org/10.1287/mnsc.2020.3678>
33. Ventura, S., Badenes-Ribera, L., Herrero, R., Cebolla, A., Galiana, L., & Baños, R. M. (2020). Virtual reality as a medium to elicit empathy: A meta-analysis. *Cyberpsychology, Behavior, and Social Networking*, 23(10), 667–676. <https://doi.org/10.1089/cyber.2019.0681>
34. Villena-Taranilla, R., Tirado-Olivares, S., Cózar-Gutiérrez, R., & González-Calero, J. A. (2022). Effects of virtual reality on learning outcomes in K-6 education: A meta-analysis. *Educational Research Review*, 35, 100434. <https://doi.org/10.1016/j.edurev.2022.100434>
35. Wu, B., Yu, X., & Gu, X. (2020). Effectiveness of immersive virtual reality using head-mounted displays on learning performance: A meta-analysis. *British Journal of Educational Technology*, 51(6), 1991–2005. <https://doi.org/10.1111/bjet.13023>
36. Zhang, Y., Zhou, H., & Li, Y. (2020). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 11, 2218. <https://doi.org/10.3389/fpsyg.2020.02218>