A comparison of threshold-free measures for assessing the effectiveness of educational interventions

Jochen Einbeck[a,c,], Tahani Coolen-Maturi[a,c], Germaine Uwimpuhwe[b,c], Akansha Singh[b,c]

[a]*Department of Mathematical Sciences, Durham University, Durham, UK*
[b]*Department of Anthropology, Durham University, Durham, UK*
[c]*Durham Research Methods Centre, Durham, UK*

Abstract

The effectiveness of educational interventions has traditionally been evaluated using effect size measures which focus on a single feature of the distribution of the outcomes under intervention and control conditions: a (standardized) mean difference. Recently there has been increased interest in methods which assess the information contained in the full distributions of outcomes under intervention and control, providing measures of separation from these distributions which do not depend on arbitrary cut-offs, hence which are "threshold-free". We investigate the statistical relationship between several concepts of this type, and discuss how they can be used to estimate alternative effect size metrics as well as their uncertainties in the context of multilevel models as commonly used for the analysis of educational data. A specific aim of this paper is to investigate how the recently proposed "gain index" relates to other measures of separation including the Area under the Curve (AUC) and the overlapping index. A simulation study, using data with an educationally motivated structure, is presented to compare the different methodologies.

Keywords: Educational trial; effect size; gain index; ROC curves; Gini coefficient; Kolmogorov-Smirnov index; Youden index

1. Introduction

In the educational sciences, the effectiveness of a new intervention (such as, a new teaching concept or a new online learning tool), is commonly investigated through experimental studies involving the randomized allocation of individuals to an intervention and a control group. The experimental units in such studies are usually pupils (possibly nested within classrooms and/or schools), and one is interested whether or not a specific educational outcome, such as a reading or writing score, has improved due to the intervention. Therefore, a post-test score (collected from pupils after delivery of the intervention) is benchmarked against a baseline score (collected from pupils before intervention delivery). Some statistical method is then applied to demonstrate whether, and to which extent, the intervention has been "effective", i.e., whether it can be evidenced to have improved the outcome of interest.

Measures of the effectiveness of interventions are most commonly based on the "effect size", which is a standardized mean difference or a standardized estimate of the intervention parameter from a fitted statistical model. While straightforward to compute, their interpretation relies on arbitrary categorizations or thresholds (such as Cohen's categorization, Cohen (1988)) or intuitive but debatable mappings (such as months of progress, Higgins et al. (2016)).

A different line of thinking attempts to assess whether individuals in the intervention group are (in some yet-to-specify sense) "more likely" to benefit from the intervention than individuals in the control group. The related statistical methods, hitherto less commonly considered in the educational context, are not based on estimated point or interval estimates of intervention effects or effect sizes, but attempt to draw inferences on the effectiveness of an intervention by making comparative distributional statements on some attainment score under the intervention and control group. An idea of this type was recently presented (Uwimpuhwe et al., 2022), aiming to estimate the proportion of pupils who actually benefitted from an intervention. Their approach uses a two-component Bayesian mixture model fitted to the "gain scores" (a term to be given more attention in the next section) in order to identify pupils as either progressed or non-progressed, which after cross-classification with the intervention assignments leads to an estimate of the proportion of pupils who progressed due to the intervention *only*.

We will show that their approach can be considered to be related to the Receiver Operating Characteristics curve (ROC), a device frequently used in reliability and medical diagnostics. Specifically, the ROC approach gives rise to a Kolmogorov-Smirnov (KS) -type statistic, as well as an estimator for the Area under the Curve (AUC) based on the Mann-Whitney U statistic. The former can be shown to have a strong conceptual link to the gain index (Uwimpuhwe et al., 2022); one can argue that both aim to measure very similar constructs. Such methodology also relates to the Overlapping coefficient (OVL) studied recently by Franco-Pereira et al. (2021). All these approaches exploit the relative positioning of two distributions, and they can do so in fully nonparametric ways, which don't require specifications of thresholds or parametric distributions. One could argue that the gain index possesses these properties as well: while it is based on a Gaussian mixture distribution, the Gaussianity of the components does not play an important role (Uwimpuhwe et al., 2022). However, the ROC and OVL-based approaches lead to statistics that are easier to compute than the gain index, and do not require neither the fitting of a mixture model nor Bayesian MCMC machinery. There are, however, nuanced differences in the notions of "assessing effectiveness" that all these statistics capture, which do not only have to do with questions of estimation or computation.

In this context, while Uwimpuhwe et al. (2022) reported a "strong correlation" between the gain index and the effect size, this assessment was based on a small sample of educational trials funded by the same organization, hence possibly with similar characteristics. Moreover, a plain correlation value does not allow any insights into whether the gain index could capture different dimensions of pupil attainment under certain scenarios, such as differing intra-class-correlations. Furthermore, little is known about the relationship of the other mentioned indexes (OVL, KS, AUC) to either effect size or gain index. Since educational trials are expensive, time-intensive, and the resulting data often difficult to access due to the requirement of using specialized analysis systems such as, in the UK, the Secure Research Service (SRS) by the Office for National Statistics, a simulation-based approach is the only practical way to comprehensively answer such questions. Therefore, in this work, we will examine, by means of a simulation study, the relationships between Uwimpuhwe et al. (2022)'s approach and these simpler probabilistic (but fully nonparametric) approaches, as described above. We will also put some emphasis on conceptual considerations linking these approaches, and on elucidating any interpretational differences between these indexes.

The paper unfolds as follows. Section 2 reviews the various threshold-free effect size measures. Specifically, Section 2.1 will recall the gain index. The several ROC-based approaches will be considered in Section 2.2, followed by a nonparametric version of the Overlapping coefficient in Section 2.3. Technical and inferential details are dealt with in Section 3. Section 4 will present the simulation study benchmarking the individual approaches, before the work is concluded in Section 5.

2. Threshold-free effect size measures

2.1. Gain Index

The gain index (Uwimpuhwe et al., 2022) aims to estimate the proportion of pupils that actually benefit from an intervention. The idea behind this approach is that, in the time between pre-and post-intervention, some pupils will make progress irrespective of the intervention. Hence, in other words, one can describe the gain index as the proportion of pupils *who would not have made progress without the intervention.*

The implementation of this idea requires a binary concept of "progress". This in turn will require a measure to assess student attainment, and a classification procedure to identify pupils as progressed or non-progressed, based on this measure.

Beginning with the former, naïvely this could be directly a post-test score of some reading or numeracy test following the intervention. However, it will usually be appropriate to account for prior attainment by adjusting for pre-test scores, either by taking differences of post- and pre-test scores, or by fitting a linear regression model of post-test versus pre-test scores, and considering the residual from the fitted model as the measure of attainment (Xiao et al., 2019). In the educational literature, the former is known as *gain score* (Zimmerman and Williams, 1982). For simplicity of presentation we follow here the terminological convention used in Uwimpuhwe et al. (2022) who refer to this measure, say y_i, as the gain score for pupil i, irrespective of whether it is obtained as a difference or as a residual. For an educational experiment, one would typically expect the pre-intervention ("baseline") distributions of scores to be very similar for intervention and control groups, but then to diverge following the intervention, rendering a stretched but not necessarily bimodal distribution of gain scores (see Figure 1 for an illustration).

*** Figure 1 about here ***

For the classification step, one fits a two-component mixture model to the joint distribution of gain scores. The mixture component corresponding to the “larger” mixture centre is associated with “progress”, and the lower component with “no progress”. Then one utilizes the mixture responsibilities, that is the posterior probabilities that each pupil corresponds to a certain mixture centre, in order to cross-classify all pupils (intervention and control) to both mixture components. This can be seen as a MAP (maximum a posteriori) estimate of one of the states “progress” or “no progress” for each pupil. Doing this for all pupils results in a 2×2 contingency table with axes intervention/control and progress/no progress, as illustrated in Table 1. For each of control and intervention group separately, one then computes the proportion of pupils who had made progress, yielding values p_1 and p_2 respectively. The gain index is the difference, $\text{GI} = p_2 - p_1$, between these two proportions.

*** Table 1 about here ***

The reader may recognize that the layout in Table 1 resembles the Binomial effect size display (BESD, Rosenthal and Rubin (1982)), one of the earlier attempts in the literature to move away from traditional effect size measures. The BESD is however quite limited in practice, as it requires prior dichotomisation of the data (if the outcome is not yet in binary form), and the computation of the “experimental success rate” derived from it makes some assumptions on the symmetry of the outcome distribution of intervention and control which can be considered questionable.

It is important to understand that, under the gain index machinery, the allocation of pupils to the progressed or non-progressed groups is entirely automated by the mixture model.

The data analyst has no control over the proportions allocated to each group. However, since only *one* mixture model is fitted to the *combined* intervention and control data, this is not a problem: If the progressed group gets larger, this will hold for both the control and the intervention arm, and increase both p_1 and p_2 accordingly, so essentially neutralising the effect of the overall mixture proportions (not "exactly" in a strict mathematical sense, but at least "in tendency").

Mathematically, the gain index can take a value between -1 and 1. However, in an education setting, it is almost impossible to obtain extreme values close to or equal to the boundaries $\{-1, 1\}$. If the intervention is effective, the proportion of progressed pupils in the intervention group (p_2) is expected to be greater than that from the control group (p_1), and this yields a positive gain index. On the other hand, a negative gain index means a greater proportion of control group pupils who have progressed ($p_1 > p_2$). Variability of the gain index estimates can be assessed via the posterior distribution yielding posterior standard errors and credible intervals.

2.2. Receiver Operating Characteristics (ROC) curves

The Receiver Operating Characteristic (ROC) is a commonly used device in reliability and medical diagnostics. It is a graphical plot that depicts the diagnostic performance of binary classifiers. This section provides a brief introduction to ROC curves and related measures, but before we delve into that, we need to introduce some notation.

Denote by Y^T the random variable producing the gain score under the intervention (T; treatment) group, and Y^C the gain score under the control group (C). Assume that we associate good progress with $Y^g > s$, for $g \in \{C, T\}$, and some positive threshold s. Then we can define the quantities

$$S_T(s) = P(Y^T > s)$$

and

$$S_C(s) = P(Y^C > s),$$

which can be seen as the analogue of the true positive rate (TPR, sensitivity) and the false positive rate (FPR, 1-specificity), respectively, as used in diagnostic test theory. The gain index could then be interpreted as an empirical estimate of $S_T(s) - S_C(s)$, for some value of s. While this value of s does not play a role in the computation of the gain index, one can think of the classification procedure induced by the mixture model as making implicitly such a choice: for most practical purposes, the MAP procedure will be monotone, i.e. for a given mixture model there will be a constant s (albeit never explicitly calculated) so that pupils with gain score larger than s will be identified as "progressed" and those with gain score smaller than s as "not progressed". The benefit of Uwimpuhwe et al. (2022)'s approach is hence clear – it avoids the explicit specification of the constant s, hence rendering the method "threshold-free" in the terminology of Yuan et al. (2015).

A similarly appealing approach to the problem is offered by Receiver Operating Characterstics (ROC) theory. Plotting $S_T(s)$ versus $S_C(s)$, over a meaningful range of constants s, produces the ROC curve, which can be explicitly described by reparametrizing $S_C(s) \equiv t$,

$$(S_C(s), S_T(s)) = (t, S_T(S_C^{-1}(t))) = (t, \mathrm{ROC}(t)), \tag{1}$$

for $t \in [0, 1]$. So, in terms of common terminology for ROC curves, the quantity Y would correspond to the *diagnostic test*, and T or C to the condition to be diagnosed. In our context, the idea is slightly different — we do not want to diagnose a condition, but we want to know whether the intervention status C or T impacts differently on the outcome Y. This is in a similar spirit to work by Pepe (2003, p.74) who found that ROC curves can be "useful in applications outside of diagnostic testing"; e.g. for "evaluating treatment effects on an outcome in a clinical trial" by capturing the separation between two distributions. The curve ROC(t) as defined in (1) is indeed meaningful for this purpose, as, clearly, effective interventions are those for which ROC(t) rises quickly.

Two well known summary statistics measuring the similarity of two distributions can be directly extracted from the ROC curve. Firstly, the Kolmogorov-Smirnov index defined as

$$\mathrm{KS} = \max_t |\mathrm{ROC}(t) - t| = \sup_{s \in (-\infty,\infty)} |P(Y^T > s) - P(Y^C > s)| = \sup_{s \in (-\infty,\infty)} |S_T(s) - S_C(s)|. \quad (2)$$

The KS index measures the maximum vertical distance between the ROC curve and the 45-degree line (uninformative test, or ineffective intervention). It takes values between 0 for the uninformative test (ineffective intervention) and 1 for the perfect intervention (Pepe, 2003). So, the KS statistics can be interpreted as a version of the gain index corresponding to the threshold s which maximizes $|S_T(s) - S_C(s)|$. It would appear plausible, considering Figure 1, that such a value s corresponds to a setting where the two distributions are well separated, suggesting that the KS statistic identifies a similar notion of "separability" of distributions as the gain index. An estimate of KS can be obtained as the maximum vertical distance between the empirical ROC curve, $\widehat{\mathrm{ROC}}$, and the 45° line, that is

$$\widehat{\mathrm{KS}} = \max_t |\widehat{\mathrm{ROC}}(t) - t| = \sup_{s \in (-\infty,\infty)} |\hat{S}_T(s) - \hat{S}_C(s)|, \quad (3)$$

with expressions for the estimates $\hat{S}_T(s)$ and $\hat{S}_C(s)$ to be given later in this subsection. The statistic $\widehat{\mathrm{KS}}$ is just the well-known Kolmogorov-Smirnov statistic for testing the equality of the two distributions. One can also use a Kolmogorov-Smirnov-based test to compare the equivalence of different ROC curves; this will not be considered further in this paper; we refer the reader to Bradley (2013).

Secondly, it is a well known fact that the integral over this curve, i.e. the *area under the curve* (AUC) is given by

$$\mathrm{AUC} = \int_0^1 \mathrm{ROC}(t)\, dt \tag{4}$$

where higher AUC values indicate more accurate tests, with AUC = 1 for perfect or ideal tests and AUC = 0.5 for uninformative tests, corresponds mathematically to the probability statement

$$\mathrm{AUC} = P(Y^T > Y^C). \tag{5}$$

That is, if one samples randomly one gain score from Y^T and one from Y^C, then (5) is the probability that $Y^T > Y^C$. Equation (5) captures directly the notion of whether an "intervention works": If it does, there should be a positive probability to the event that $Y^T - Y^C > 0$. An idea of this type was also pursued by McGraw and Wong (1992), approximating expression (5) based on normality assumptions on Y^C and Y^T. We will calculate (5) without any distributional assumptions.

One could of course be more stringent and demand that $Y^T - Y^C > \delta$, for some educationally relevant threshold $\delta > 0$. This would however bring back the requirement to specify that threshold, which we have been setting out to avoid.

One could think of this AUC-based concept as integrating the performance of the diagnostic test across all possible thresholds, hence avoiding the need for its choice. The conceptual analogy between (4) and (5) suggests that AUC can be used to make a statement on whether the intervention has worked, in the sense of: *If the gain score was used as a diagnostic test* *between the intervention and control 'condition', then how well would it discriminate?*

Of course, implementing this concept in practice will require us to be able to estimate (5) from real data. Denote by y_i^T, $i = 1, \ldots, n$ and y_j^C, $j = 1, \ldots, m$ the measured gain scores relating to the intervention and control group, respectively. Firstly we note that both $S_T(s)$ and $S_C(s)$ are in principle easily estimable through the empirical survivor functions $\hat{S}_T(s) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{y_i^T > s\}$ and $\hat{S}_C(s) = \frac{1}{m}\sum_{j=1}^{m} \mathbf{1}\{y_j^C > s\}$, where $\mathbf{1}\{E\}$ is an indicator function which is equal to 1 if event E occurs and 0 else. These estimates could then be used to build an empirical ROC curve $\{(\hat{S}_C(s), \hat{S}_T(s)),\ s \in (-\infty, \infty)\}$, which then would lead to the empirical AUC through stepwise integration (i.e. summation). However, this is actually not necessary as one can estimate AUC in a much simpler way based on the Mann-Whitney U test known from nonparametric test theory (Pepe, 2003). This is a test for the null hypothesis that, for randomly selected values A and B from two populations, the probability of A being greater than B is equal to the probability of B being greater than A. It can be seen as a test that compares whether the distribution of a dependent variable is the same for two groups, and therefore whether the two groups belong to the same population. We now identify A notationally with Y^T and B with Y^C. Then the corresponding test statistic, known as the Mann-Whitney U statistic, is given by

$$U = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[\mathbf{1}\{y_i^T > y_j^C\} + \frac{1}{2}\,\mathbf{1}\{y_i^T = y_j^C\}\right]. \tag{6}$$

One can then show (Hanley and McNeil, 1982) that the empirical area under the curve corresponds just to

$$\widehat{\mathrm{AUC}} = \frac{U}{nm}, \tag{7}$$

hence opening up a quick computational device for computing the AUC, and so for the computation the empirical estimate of (5). Further ROC summary indices do exist, such as the partial area under ROC (Pepe, 2003), which are not considered further in this work.

2.3. The Overlapping Index (OVL)

The overlap coefficient (OVL) quantifies the similarity (or difference) between two distributions via the overlapping area of their probability density functions, e.g. the two plots in Figure 2 are examples of low and high overlapping probability densities.

*** Figure 2 about here ***

Now let f_{Y^T} and f_{Y^C} be the corresponding probability densities (of some gain score), for the intervention and control groups, respectively. The overlap coefficient is the overlap area between the two densities, defined as

$$\mathrm{OVL} = \int \min[f_{Y^T}(y), f_{Y^C}(y)]dy. \tag{8}$$

The index OVL takes a value between 0 and 1, where OVL $= 0$ if the two probability densities are disjoint and OVL $= 1$ if the two densities are identical. There are no clear cutoff values to describe the discrimination ability of this measure, but a rule of thumb has been suggested (Franco-Pereira et al., 2021) as follows:

$$\begin{cases} \mathrm{OVL} = 1 & \text{no differentiation;} \\ 0.75 < \mathrm{OVL} < 1 & \text{poor differentiation;} \\ 0.55 < \mathrm{OVL} < 0.75 & \text{good differentiation;} \\ 0.35 < \mathrm{OVL} < 0.55 & \text{very good differentiation;} \\ \mathrm{OVL} < 0.35 & \text{excellent differentiation.} \end{cases}$$

Several fully nonparametric, kernel-based approaches for the computation of the overlap coefficient have been introduced and studied in the literature (Pastore and Calcagnì, 2019;

Franco-Pereira et al., 2021). Thus, the densities in Equation (8) are replaced by appropriate kernel density estimators. The density estimator for $f_{Y^T}(y)$ is given by

$$\hat{f}_{Y^T}(y) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{y - y_i^T}{h}\right)$$

where K is a Gaussian kernel function and h a bandwidth parameter which can can be automatically selected by

$$h = (4/3)^{1/5} n^{-1/5} q, \quad \text{where } q = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i^T - \sum_{j=1}^{n}\frac{y_j^T}{n}\right)^2}.$$

Similarly we can define $\hat{f}_{Y^C}(y)$, and substituting these densities estimates in (8) yields the following estimate for the overlapping index:

$$\widehat{\text{OVL}} = \int \min[\hat{f}_{Y^T}(y), \hat{f}_{Y^C}(y)]dy. \tag{9}$$

3. Technical and inferential aspects

3.1. Multilevel adjustment

The gain index approach uses a Bayesian "shared parameter mixture model" (Evans and Erlandson, 2004) in order to account for the within-class correlation of a potential multilevel structure in the data. It does not appear obvious how such an approach could be followed in the context of either the ROC-based measures or the OVL statistic. However, a simple way around this problem is to use the residuals of an appropriate multilevel model (accounting for pre-test but excluding the intervention effect), rather than of a plain regression model, in order to compute the gain scores. We will apply this approach for the KS, Gini (AUC), and OVL measure in Section 4.

Of course, one could then in principle also use this approach for the gain index itself. It would be expected that the resulting quantity behaves similarly to the original gain index (GI). Hence, for comparison purposes, a new version of the gain index to which we refer as "residual gain index" (GI_{res}), that uses the same residual information as ROC (and its derived quantities) or the OVL statistics is considered along with the gain index (GI). The GI_{res} accounts for clustering, at the first step of the gain index computation, by using a multilevel model to obtain the residuals or gain scores, and an unclustered finite mixture model in the second step. In contrast, the original GI obtains residuals from the OLS model, with clustering being accounted for in the second step using a finite mixture model.

3.2. Use of indexes as effect size measures

Several of the indexes mentioned in Sections 2.2 and 2.3 are closely related, or even equivalent, under some assumptions or certain transformations. Understanding these de-facto equivalences is important when carrying out any quantitative study of these metrics, as we aim to do in Section 4. Therefore, we summarize these relationships here.

Firstly, we observe that, for the KS statistic (3), large values correspond to large separations of the two involved distributions, and that this statistic will always reside between 0 (no separation) and 1 (maximal separation). A closely related statistic is the Youden Index, $J = \sup_{s\in(-\infty,\infty)}(S_T(s) - S_C(s))$, which is the maximum distance between the true positive and false positive rates (Krzanowski and Hand, 2009). Since the Youden index assumes that the proportion of positive results for the intervention is greater or equal than that for the control (Youden, 1950), i.e.

$$S_T(s) \geq S_C(s) \quad \text{for all } \; s, \tag{10}$$

the two quantities KS and J are equivalent under this assumption. However, when the intervention is less effective than the control, then J will be undefined but KS will still give an undirected measure of the separability of the distributions. The property (10) is also known as *stochastic dominance* (Martínez-Camblor, 2023) and implies non-crossing ROC curves.

Secondly, we also know from the Proposition in Section 3 of Martínez-Camblor (2023) that, again under the assumption of stochastic dominance, the (distributional version of the) overlapping index is just given by 1 minus the Youden index. Hence, for the sake of general comparability, we will consider the quantity $1 - \text{OVL}$, rather than the overlapping index OVL by itself, in our comparative analysis. The quantity $1 - \text{OVL}$ takes then the value 0 for complete identity of distributions and 1 for maximum separability.

Thirdly, a notable variant of the AUC is the Gini coefficient, $2 \times \text{AUC} - 1$, which has some benefits in terms of interpretability as compared to AUC itself. Specifically, while the AUC ranges from 0 to 1, with the value 1/2 meaning no separation, the Gini coefficient ranges from -1 to 1, with 0 representing no separation, in line with the gain index.

Summarizing, the gain index GI, its residual variant GI_{res}, and the Gini coefficient, take values between -1 and 1, and the statistics KS and $1 - \text{OVL}$ between 0 and 1, where however 0 always means "no separation" under all statistics. Furthermore, complete separation of the intervention and control group, in favor of the control, will for all of these statistics correspond to the value of 1, so that for practical purposes the statistics can be considered to be on equal scale, and hence can be sensibly compared in the range [0,1]. A comparison of the ranges and properties of these statistics is provided in Table 2, and a summary display putting these criteria in conceptual relation with each other is provided in Figure 3. We give a quantitive comparison of these measures in the simulation study in Section 4.

*** Table 2 about here ***

*** Figure 3 about here ***

It is noted finally that theoretical equivalences of the type just mentioned do not necessarily translate into exactly matching numerical results in practical experiments. This is because the *estimators* of these quantities are partially based on different principles; for instance, as outlined in Section 2.3, for the overlapping index, some smoothing is involved, while for the other statistics it isn't. This explains that, while under stochastic dominance one has in theory (as argued above)

$$\mathrm{KS} = J = 1 - \mathrm{OVL},$$

the *estimates* $\widehat{\mathrm{KS}}$ and $1 - \widehat{\mathrm{OVL}}$ will not usually coincide, even if stochastic dominance is fulfilled.

3.3. Estimation and uncertainty

In practice, there is the important question of which software or tool to use in order to practically estimate the quantities elaborated on. Furthermore, in order to make robust decisions on the effectiveness of an intervention, it is essential that one is able to quantify the uncertainty of the measures considered. In this section we summarize how this is achieved for the metrics under consideration.

The estimation of the gain index (including its residual variant) is intrinsically Bayesian. Hence, the estimated gain index is obtained as the mean from its posterior distribution, and estimates of its uncertainty can be naturally obtained from appropriate quantiles of this posterior distribution. In practice, we obtain gain index estimates and their Bayesian credible intervals using the r2jags package (Su et al., 2015) and the function provided in the supplementary material of Uwimpuhwe et al. (2022).

Estimates of KS were obtained using the stats package, and a bootstrapping procedure with 1000 iterations was employed to determine their 95% confidence intervals.

Estimates and confidence intervals for $\widehat{\mathrm{AUC}}$, and by extension for the estimate of the Gini index, $\widehat{\mathrm{Gini}} = 2 \times \widehat{\mathrm{AUC}} - 1$, are obtained using the pAUC package (Robin et al., 2011).

Estimates of the OVL have been obtained using the kernel method described earlier Franco-Pereira et al. (2021), and the variance of this estimator is estimated using bootstrap, enabling the computation of confidence intervals for OVL (Pastore and Calcagnì, 2019; Franco-Pereira et al., 2021).

4. Simulation study

4.1. Simulation model setup

In order to illustrate and compare the methods presented in this paper in an educationally relevant setting, we simulated data from a cluster randomized trial (CRT) design, which is a commonly used design for trials commissioned by the Education Endowment Foundation (EEF). To account for CRT design, the following model suggested in the EEF statistical guidelines (Education Endowment Foundation, 2022), which controls for baseline measurements and accounts for school clustering, was considered:

$$Y_{ij} = \beta_0 + \beta_1 P_{ij} + \beta_2 T_{ij} + b_i + \epsilon_{ij}. \tag{11}$$

Here, Y_{ij} and P_{ij} are respectively the post- and pre-intervention outcomes of pupil i from school j, T_{ij} is a two-arms intervention indicating whether pupil i from school j was randomized to receive control (reference category; $T_{ij} = 0$) or intervention ($T_{ij} = 1$), $b_i \sim N(0, \sigma_b^2)$

is a school-specific random intercept, with σ_b capturing between-school variability, and $\epsilon_{ij} \sim N(0, \sigma^2)$ is pupil-level random error, with σ capturing within-school variability.

In order to simulate data that capture two natural groupings of educational outcomes, we included a latent groups variable (LG) indicating progress ($LG_{ij} = 1$) and no progress (reference category; $LG_{ij} = 0$) into the model given by Equation (11). This led us to consider the following simulation model:

$$Y_{ij} = \beta_0 + \beta_1 P_{ij} + \beta_2 T_{ij} + \beta_3 LG_{ij} + b_i + \epsilon_{ij}. \tag{12}$$

The second parameter β_3 controls the extent to which the average post-test attainment between the LG groups differs, or in other words, the degree of separation between the progressed and non-progressed groups. The degree of separation incurred by this process is illustrated in Figures 4 and 5, for the exemplary settings of $\beta_3 = 2, 4,$ and 8, through overlapping plots and ROC curves, respectively, using a randomly generated data set from each of the three settings. We consider these as to be corresponding to small, medium, and large separation. We see that, while the OVL representation distinguishes the three cases very clearly, in the ROC representation the distinction between the three possible settings appears less obvious, which would lead us to suspect that OVL is more sensitive to β_3 than the AUC-based measures (KS and Gini) are.

*** Figure 4 about here ***

*** Figure 5 about here ***

4.2. Separation measures as a function of effect size and latent group separation

We initially fix the residual variance $\sigma^2 = 1^2$ and the cluster variance $\sigma_b^2 = 0.34^2$, implying an intra-class correlation (ICC) of approximately 0.1 as typically encountered for educational interventions in the UK (Singh et al., 2023). In this simulation study we have considered fifteen scenarios to simulate data sets from, based on five choices of β_2 , and three choices of β_3. Specifically, we determine the five values of β_2 using the fact that $\beta_2 = \text{ES} \times \sqrt{\sigma^2 + \sigma_b^2}$, with the effect size settings $\text{ES} = -0.15, 0.05, 0.2, 0.5, 0.8$, the last three of which corresponding to small, medium and large effect sizes according to Cohen (1988). The values -0.15 and 0.05 correspond to negative and very small effect sizes, which are not uncommon in educational settings (Ashraf et al., 2021). For β_3, we consider the settings 2, 4 and 8 as motivated in the previous subsection.

The LG_{ij} and T_{ij} variables were simulated so that the percentage of progressed pupils ($LG_{ij} = \text{progress}$) was 0.462 in the control group ($T_{ij} = 0$) and 0.750 in the intervention group ($T_{ij} = 1$), resulting in a "true" GI of $0.750 - 0.462 = 0.288$. Consequently, the same true GI would be obtained for each simulated data by cross-tabulating the T_{ij} and LG_{ij} variables. For each of the two treatment arms (control and intervention), we simulated eight schools, each with a size of 30 pupils. This yields a total sample size of 480 for each dataset. The baseline scores (P_{ij}) were generated from $N(0, 0.5^2)$, random schools (b_i) from $N(0, 0.34^2)$ and residuals (ϵ_{ij}) from $N(0, 1^2)$. The outcome (Y_{ij}) was obtained from equation (12), where β_0 was fixed to 0 and β_1 to 0.5.

For each of the fifteen scenarios, 100 datasets were simulated. The average estimates across scenarios for GI, GI_{res}, $\widehat{\text{KS}}$, $\widehat{\text{Gini}}$ and $1 - \widehat{\text{OVL}}$, are shown in Table 3. As specified in Section 3.1, the gain scores (Y_{ij}) used in calculation of $\widehat{\text{KS}}$, $\widehat{\text{Gini}}$, $1 - \widehat{\text{OVL}}$ and GI_{res} were defined as residuals (ϵ_{ij}) from a fitted multilevel model of type (11) but with intervention terms ($\beta_2 T_{ij}$) excluded. The Y_{ij} scores were subsequently organized based on intervention groups to obtain Y_{ij}^g representing gain scores for the intervention group (i.e. $g = T$ if $T_{ij} = 1$)

and Y_{ij}^C for the control group (i.e. $g = C$ if $T_{ij} = 0$), respectively.

*** Table 3 about here ***

Several observations can be drawn from Table 3. First, we find that all measures tend to increase with the true effect size, for all settings of β_3. Secondly, we find that the "true gain index" of 0.288 is reasonably well estimated by both the original and the residual version of the gain index, with estimation getting more precise for larger effect sizes and larger separation between the progressed and non-progressed groups. We also find that the original and the residual-based gain index estimates can hardly be distinguished; their values are very similar, and in fact very slightly smaller for GI_{res} than for GI. The confidence bands of the residual version appear however to be slightly wider than those of the original gain index version, indicating a practical advantage of the latter. Thirdly, we find that $\widehat{\text{KS}}$ and $1-\widehat{\text{OVL}}$ behave broadly similarly as a function of ES, but the Gini measure covers a larger range of the unity interval, in fact "overtaking" the other measures on the way from small to large effect sizes. On the other hand, but as expected from previous considerations, $1-\widehat{\text{OVL}}$ and the gain index show a stronger dependency on the parameter β_3 than $\widehat{\text{KS}}$ and $\widehat{\text{Gini}}$; specifically their performance at identifying the intervention effect increases when the attainment difference between the progress groups is generally larger.

*** Figure 6 about here ***

A more illuminating view on the results is provided through a graphical representation. Figure 6 shows the five threshold-free measures against the true effect size, for $\beta_3 = 2, 4$ and 8. We see from the panels in that figure that, perhaps rather surprisingly, all of the considered threshold-free, probabilistic separation measures are almost perfectly linearly related to the true underlying effect size. Specifically, all measures, except Gini (AUC), show largely parallel behavior, that is they capture the same information on distributional separability, but subject to some vertical offset. In particular, this confirms our previous theoretical considerations that KS and $1-\mathrm{OVL}$ provide equivalent information, which is in turn very strongly related to the one provided by the gain index. All these criteria provide information on the proportion of pupils benefitting from an intervention, rather than on the size of the effect. However, only the gain index allows *exactly* this interpretation.

One can also observe that these curves get flatter once the parameter β_3 increases. For the gain index, this can be intuitively explained: After all, the "true" gain index, under all simulation scenarios, is constant at 0.288. This means that the gain index shows the less bias, the larger the underlying separation between progressed and non-progressed groups; a behavior which makes fully sense given the methodological approach behind this index. In other terms, smaller slopes in Figure 6 indicate that a dimension of progress closer to the gain index is captured by the respective measure, whereas larger slopes mean that overall progress, as induced by the simulated ES, is measured. If the underlying gap between progressed and non-progressed pupils gets larger (β_3 increases), it becomes harder for the intervention to move pupils from the non-progressed to the progressed group (even if they do make some progress), hence the traditional effect size and the gain index become less correlated.

The behavior of these four criteria is different to that of the Gini coefficient (i.e, AUC), which shows a steeper behavior than the other measures, for any choice of the progress separation parameter β_3. Hence, AUC addresses a dimension of pupil achievement which is more closely related to the original effect size index. Also, we observe that the only other measure which still enables some effect size–dependent discrimination for highly separated latent progress groups is the KS – which is not unsurprising given that it shares with the

AUC the conceptual reliance on ROC curves.

4.3. Separation measures as a function of ICC, pre-test, and true gain index

In this subsection we provide some additional simulation results which show the impact of some other auxiliary simulation settings. While these add limited insight to the relationship of the indices between each other, they do shed light on the robustness of the results w.r.t. deviations from the parameter settings of the simulation in the previous subsection. Firstly, noting that the "true" gain index of 0.288 is rather large in comparison with values typically found for educational trials, we consider additional settings of 0.05 and 0.12, corresponding in magnitude to the results reported in Uwimpuhwe et al. (2022). Secondly, we consider a larger ICC of 0.2, resulting from a school-level variance of $\sigma_b^2 = 0.5$. Such an ICC value is still consistent with values reported in the educational literature (Zopluoğlu, 2012). Thirdly, we also compare two settings of the pre-test parameter β_1. The impact of baseline effects on effect size estimates is generally expected to be small (Verbeke and Fieuws, 2007); hence it will be interesting to see whether this also holds for the various separation measures considered. In these simulations, we leave the progress group separation parameter $\beta_4 = 4$ and the effect size $ES = 0.2$ fixed, corresponding to intermediate values used in Table 3. Finally, in this framework, we consider the additional settings of ICC = 0.3 and ICC = 0.4 when the true gain index is 0.288 and $\beta_1 = 1$. While such high ICC values are unlikely to be observed in educational trials, the results may shed light on the theoretical behavior of the criteria when within-school correlations are extremely large.

*** Table 4 about here ***

Results of this simulation study are reported in Table 4. Focusing initially on the results above the dashed line, we observe that the gain index is again reasonably estimated under all settings, where however the smaller gain index settings appear to be easier estimable than the highest one. The difference between GI and GI$_{res}$ also increases as the true gain index increases, with the values being closer when ICC is smaller. All displayed criteria do replicate the design pattern imposed by the true gain index. All separation measures do show some slight sensitivity to the ICC (i.e. σ_b^2), with the sensitivity being smallest for the gain index estimation itself. As expected, the pre-test parameter β_1 does not play any role, apart from a very slight impact on the overlapping coefficient. Considering now the results for the higher ICC settings of 0.3 and 0.4, we observe that all measures, except the gain index itself, show some weakness in maintaining the levels of separation identified at ICC $= 0.1$ and ICC $= 0.2$. This appears to indicate that the multilevel adjustment through the Bayesian mixture model, as used in the gain index, is more effective than the residual-based adjustment used for all other indicators (including the residual gain index). It is mentioned however again that ICC values of 0.3 and higher are unlikely to actually occur in educational trials.

5. Discussion

Through conceptual considerations and simulations, we have illustrated that the gain index is strongly related, and in fact measures equivalent information, to the Overlapping index and, taking some slightly different behavior under rather extreme scenarios aside, also to the KS index. On the other hand, the AUC criterion reflects a quantity that is more related to the original effect size measure. Whether or not these two families of measures correlate strongly depends on the gap in the latent distribution of progressed and non-progressed students: If that gap is small, then the two families of measures will be linearly related, with the gain index possibly providing more useful information, and indicating progress of a *subgroup of pupils* even if the overall effect size is close to zero, or in fact negative. However, the gain index will measure a different dimension of student progress than the original effect size (typically estimates by Hedges' g) and Gini (AUC) when the underlying progress gap is large, as in this case it is less likely for the intervention to 'flip' a pupil over. An interesting intermediate position in this respect is taken by the KS statistic, which behaves similarly to the gain index for small separation of progress groups, but continues to measure some aspect of overall (mean) progress under large separation.

Summarizing, for most applied scenarios, considering *any of* gain index, residual gain index, or the overlapping index, in addition to AUC or Gini (or Hedges' g), will provide a pretty complete insight into the student progress accomplished by an educational intervention: Large values of the former indicate whether the intervention has benefitted a large proportion of pupils; large values of the latter indicate whether there has been a good overall (mean) progress across pupils. While the KS statistic appears to be in the unique position of capturing elements of the two dimensions at the same time, it is still stronger related to the gain index family than to traditional effect sizes. When choosing between the gain index, the OVL, or KS, then computational simplicity, the availability of many standard software packages for their computation, the solid grounding in the statistical literature, and perhaps a slightly better comprehensibility of the statistics themselves, speak for the latter two measures. If the focus is on *relative comparisons* of models, parameter configurations, or data sets, then OVL or KS will fully serve this purpose.

However, the gain index may still have an advantage in terms of its interpretability, as it is designed to measure *directly* what the educational researcher may be interested in: the proportion of pupils that have benefitted from an intervention. This may outweigh the more elaborated computations required and the perhaps additional effort required by the researcher to understand that it does exactly this.

As with all simulation work, a word of caution is in order, in the sense that our results are valid within the ranges of the parameter settings explored in our simulation studies. We have given a glimpse of this limitation by considering, in subsection 4.1, ICC values of 0.3 and 0.4, despite solid evidence from both US and UK contexts that such high ICC values are extremely rare to occur in educational trials (Hedges and Hedberg (2007), Singh et al. (2023)). Hence, further experiments to explore other simulation settings are still encouraged. Some additional interest could, for instance, lie in "degenerate" or boundary cases, only partially explored in here, such as where the gain index or progress group separation approach zero. Future studies could also focus on the application and validation of threshold-free measures in diverse real world educational settings, as these measures have not been widely used in applied research. This would help in understanding their effectiveness and limitations in real-world conditions.

Presenting simple and interpretable measures of educational intervention effectiveness is crucial for education policy-making (Gorard et al., 2020) as it provides a clear and concise way to assess the impact of interventions on educational outcomes. This is important to make informed decisions based on evidence. Research indicates that educational interventions can vary widely in their design, delivery, and outcomes (Buhl-Wiggers et al., 2022). Further, since pupil's educational outcomes vary a lot, a holistic evaluation of the effectiveness of interventions requires the availability of measures which capture the information in the full outcome distributions (Uwimpuhwe et al., 2022) rather than focusing on mean differences as in traditional effect size measures, which help assessing program effectiveness at a broader level (Higgins et al., 2016).

The threshold-free (non-parametric, probabilistic) measures of separation enhance our understanding of educational interventions by considering the entire outcome distribution, thereby also capturing individual student progress. Educators and researchers should embrace advantages of these measures and use them alongside traditional metrics to inform evidence-based decision-making. Utilising these simple and interpretable measures, educational policy makers and school teachers can easily compare different interventions and determine which ones are most effective in improving educational outcomes. This streamlined approach would help in allocating resources efficiently and implementing evidence-based policies that have a positive impact on education systems.

Acknowledgements

(removed from blinded version)

Declaration of interest statement

The authors report there are no competing interests to declare.

References

Ashraf, B., Singh, A., Uwimpuhwe, G., Higgins, S., and Kasim, A. (2021). Individual participant data meta-analysis of the impact of educational interventions on pupils eligible for free school meals. *British Educational Research Journal*, 47(6):1675–1699.

Bradley, A. P. (2013). ROC curve equivalence using the Kolmogorov–Smirnov test. *Pattern Recognition Letters*, 34(5):470–475.

Buhl-Wiggers, J., Kerwin, J. T., Muñoz-Morales, J., Smith, J., and Thornton, R. (2022). Some children left behind: Variation in the effects of an educational intervention. *Journal of Econometrics*.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)*.

Education Endowment Foundation (2022). Statistical analysis guidance for EEF evaluations.

Evans, R. and Erlandson, K. (2004). Robust bayesian prediction of subject disease status and population prevalence using several similar diagnostic tests. *Statistics in Medicine*, 23(24):2227–36.

Franco-Pereira, A. M., Nakas, C. T., Reiser, B., and Carmen Pardo, M. (2021). Inference on the overlap coefficient: The binormal approach and alternatives. *Statistical Methods in Medical Research*, 30(12):2672–2684.

Gorard, S., See, B. H., and Siddiqui, N. (2020). What is the evidence on the best way to get evidence into use in education? *Review of Education*, 8(2):570–610.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.

Hedges, L. V. and Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29:60–87.

Higgins, S., Katsipataki, M., Villanueva-Aguilera, A., Coleman, R., Henderson, P., Major, L., Coe, R., and Mason, D. (2016). *The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit, Manual.* Education Endowment Foundation, London.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC Curves for Continuous Data.*

Martínez-Camblor, P. (2023). About the use of the overlap coefficient in the binary classification context. *Communications in Statistics - Theory and Methods*, 52(19):6767–6777.

McGraw, K. O. and Wong, S. P. (1992). *Psychological Bulletin*, 111:361–365.

Pastore, M. and Calcagnì, A. (2019). Measuring distribution similarities between samples: A distribution-free overlapping index. *Frontiers in Psychology*, 10:1089.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* Oxford University Press, Oxford.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, 12:77.

Rosenthal, R. and Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology*, 74:166—-169.

Singh, A., Uwimpuhwe, G., Vallis, D., Akhter, N., Coolen-Maturi, T., Higgins, S., Einbeck, J., Culliney, M., and Demack, S. (2023). Improving power calculations in educational trials.

Su, Y.-S., Yajima, M., Su, M. Y.-S., and SystemRequirements, J. (2015). Package ‘r2jags’. *R package version 0.03-08, URL http://CRAN. R-project. org/package= R2jags.*

Uwimpuhwe, G., Singh, A., Higgins, S., Coux, M., Xiao, Z., Shkedy, Z., and Kasim, A.

(2022). Latent class evaluation in educational trials: What percentage of children benefits from an intervention? *The Journal of Experimental Education*, 90(2):404–418.

Verbeke, G. and Fieuws, S. (2007). The effect of miss-specified baseline characteristics on inference for longitudinal trends in linear mixed models. *Biostatistics*, 8(4):772–783.

Xiao, Z., Higgins, S., and Kasim, A. (2019). An empirical unraveling of lord's paradox. *The Journal of Experimental Education*, 87(1):17–32.

Youden, W. (1950). Index for rating diagnostic tests. *Cancer*, 1(3):32–35.

Yuan, Y., Su, W., and Zhu, M. (2015). Threshold-free measures for assessing the performance of medical screening tests. *Frontiers in Public Health*, 3.

Zimmerman, D. W. and Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement*, pages 149–154.

Zopluoğlu, C. (2012). A cross-national comparison of intra-class correlation coefficient in educational achievement outcomes. *Journal of Measurement and Evaluation in Education and Psychology*, 3(1):242–278.

Table 1: Computation of gain index from pupil numbers cross-classified to progressed and non-progressed groups. Adapted from Table 2 in Uwimpuhwe et al. (2022).

Random assignment	non-progressed	progressed	proportion	gain index
Control	a	b	$p_1 = b/(a+b)$	
Intervention	c	d	$p_2 = c/(c+d)$	$\text{GI} = p_2 - p_1$

Table 2: Separation and effect size measures with their operational ranges, as well as key values of no and perfect separation. For the U statistic, n and m indicate the sample size of the control and intervention group, respectively. Where these are theoretical quantities which need estimation (for KS, AUC, Gini, J, and OVL), the properties continue to hold true for the estimates $\widehat{\text{KS}}$, $\widehat{\text{AUC}}$, $\hat{J}$, $\widehat{\text{Gini}}$, $\widehat{\text{OVL}}$. The two gain index measures, as well as U, are by construction empirical quantities without distributional counterparts.

			perfect separation in favour of...	
Criterion	range	no separation	control	intervention
GI	[-1,1]	0	-1	1
GI$_{res}$	[-1,1]	0	-1	1
KS	[0,1]	0	1	1
J	[0,1]	0	NA	1
AUC	[0,1]	1/2	0	1
U	$[0, nm]$	$\frac{nm}{2}$	0	nm
Gini	[-1,1]	0	-1	1
OVL	[0,1]	1	0	0
$1 - \text{OVL}$	[0,1]	0	1	1

Table 3: Results of the simulation study with 0.288 as the true gain index. $\widehat{\text{Gini}}$ corresponds to $2 \times (\widehat{\text{AUC}} - 0.5)$. Each number in the table gives an average value over 100 simulated data sets. 95% CIs are given between brackets.

Scenario:	GI	GI_{res}	$\widehat{\text{KS}}$	$\widehat{\text{Gini}}$	$1 - \widehat{\text{OVL}}$
$\beta_3 = 2$					
ES $= -0.15$	0.117 [0.053, 0.180]	0.106 [0.039, 0.171]	0.178 [0.120, 0.274]	0.178 [0.076, 0.278]	0.151 [0.100, 0.229]
ES $= 0.05$	0.157 [0.093, 0.219]	0.145 [0.078, 0.210]	0.228 [0.167, 0.323]	0.254 [0.156, 0.354]	0.196 [0.135, 0.273]
ES $= 0.2$	0.186 [0.122, 0.249]	0.173 [0.106, 0.238]	0.265 [0.204, 0.359]	0.312 [0.214, 0.408]	0.231 [0.166, 0.307]
ES $= 0.5$	0.243 [0.178, 0.306]	0.229 [0.161, 0.294]	0.337 [0.277, 0.430]	0.418 [0.326, 0.510]	0.301 [0.234, 0.375]
ES $= 0.8$	0.298 [0.231, 0.362]	0.283 [0.213, 0.349]	0.409 [0.349, 0.498]	0.516 [0.432, 0.602]	0.370 [0.303, 0.441]
$\beta_3 = 4$					
ES $= -0.15$	0.233 [0.191, 0.273]	0.226 [0.182, 0.269]	0.274 [0.202, 0.364]	0.246 [0.146, 0.348]	0.252 [0.186, 0.333]
ES $= 0.05$	0.249 [0.208, 0.289]	0.243 [0.200, 0.286]	0.291 [0.223, 0.381]	0.302 [0.202, 0.400]	0.269 [0.200, 0.350]
ES $= 0.2$	0.261 [0.221, 0.301]	0.256 [0.213, 0.298]	0.305 [0.240, 0.396]	0.342 [0.246, 0.438]	0.283 [0.213, 0.363]
ES $= 0.5$	0.284 [0.243, 0.323]	0.280 [0.237, 0.322]	0.340 [0.279, 0.432]	0.422 [0.330, 0.514]	0.314 [0.246, 0.392]
ES $= 0.8$	0.306 [0.265, 0.346]	0.304 [0.260, 0.346]	0.388 [0.326, 0.477]	0.496 [0.410, 0.582]	0.350 [0.284, 0.425]
$\beta_3 = 8$					
ES $= -0.15$	0.282 [0.264, 0.298]	0.281 [0.263, 0.298]	0.293 [0.218, 0.380]	0.250 [0.150, 0.350]	0.319 [0.249, 0.405]
ES $= 0.05$	0.283 [0.266, 0.299]	0.283 [0.265, 0.299]	0.300 [0.229, 0.388]	0.302 [0.204, 0.402]	0.318 [0.247, 0.404]
ES $= 0.2$	0.285 [0.268, 0.301]	0.284 [0.266, 0.301]	0.307 [0.242, 0.398]	0.342 [0.246, 0.438]	0.318 [0.248, 0.403]
ES $= 0.5$	0.287 [0.270, 0.304]	0.287 [0.269, 0.304]	0.338 [0.277, 0.431]	0.420 [0.328, 0.512]	0.323 [0.258, 0.404]
ES $= 0.8$	0.290 [0.273, 0.308]	0.290 [0.271, 0.308]	0.387 [0.324, 0.475]	0.492 [0.406, 0.578]	0.340 [0.280, 0.414]

Table 4: Results of the simulation study with the true gain index $\mathrm{GI}_{true} = 0.05, 0.120, 0.288$; intra-class correlation ICC $= 0.1, 0.2$ (corresponding to $\sigma_b^2 = 0.34, 0.5$); and pre-test parameters $\beta_1 = 0.5, 1$; with fixed $\beta_3 = 4$; and ES $= 0.2$. $\widehat{\mathrm{Gini}}$ corresponds to $2 \times (\widehat{\mathrm{AUC}} - 0.5)$. Each number in the table gives an average value over 100 simulated data sets. 95% CIs are given between brackets. Additional simulation results for ICC $= 0.3$ and ICC $= 0.4$ are given below the dashed line.

Scenario:							
GI_{true}	ICC	β_1	GI	GI_{res}	$\widehat{\mathrm{KS}}$	$\widehat{\mathrm{Gini}}$	$1 - \widehat{\mathrm{OVL}}$
0.050	0.10	0.50	0.058 [0.017, 0.099]	0.059 [0.016, 0.102]	0.119 [0.083, 0.214]	0.100 [-0.004, 0.202]	0.104 [0.071, 0.184]
0.050	0.10	1.00	0.058 [0.017, 0.099]	0.059 [0.016, 0.102]	0.119 [0.083, 0.214]	0.100 [-0.004, 0.202]	0.104 [0.071, 0.183]
0.050	0.20	0.50	0.059 [0.018, 0.100]	0.059 [0.009, 0.108]	0.117 [0.082, 0.212]	0.098 [-0.006, 0.200]	0.101 [0.069, 0.181]
0.050	0.20	1.00	0.059 [0.018, 0.100]	0.059 [0.009, 0.108]	0.117 [0.082, 0.212]	0.098 [-0.006, 0.200]	0.101 [0.069, 0.181]
0.120	0.10	0.50	0.119 [0.078, 0.160]	0.117 [0.074, 0.161]	0.165 [0.112, 0.259]	0.172 [0.070, 0.274]	0.143 [0.094, 0.227]
0.120	0.10	1.00	0.119 [0.078, 0.160]	0.117 [0.074, 0.161]	0.165 [0.112, 0.260]	0.172 [0.070, 0.274]	0.146 [0.095 ,0.229]
0.120	0.20	0.50	0.120 [0.079, 0.161]	0.111 [0.061, 0.160]	0.161 [0.109, 0.257]	0.166 [0.064, 0.270]	0.138 [0.090, 0.220]
0.120	0.20	1.00	0.120 [0.079, 0.161]	0.111 [0.061, 0.160]	0.161 [0.109, 0.257]	0.166 [0.064, 0.270]	0.138 [0.090 ,0.220]
0.288	0.10	0.50	0.261 [0.221, 0.301]	0.256 [0.213, 0.298]	0.305 [0.239, 0.396]	0.342 [0.246, 0.438]	0.282 [0.213, 0.362]
0.288	0.10	1.00	0.261 [0.221, 0.301]	0.256 [0.213, 0.298]	0.305 [0.239, 0.396]	0.342 [0.246, 0.438]	0.283 [0.215, 0.363]
0.288	0.20	0.50	0.262 [0.221, 0.301]	0.238 [0.188, 0.286]	0.299 [0.235, 0.390]	0.336 [0.240, 0.434]	0.272 [0.205, 0.350]
0.288	0.20	1.00	0.262 [0.221, 0.301]	0.238 [0.188, 0.286]	0.299 [0.235, 0.390]	0.336 [0.240, 0.434]	0.275 [0.207, 0.352]
0.288	0.30	1.00	0.260 [0.220, 0.300]	0.213 [0.158, 0.267]	0.283 [0.220, 0.376]	0.322 [0.225, 0.420]	0.254 [0.187, 0.331]
0.288	0.40	1.00	0.262 [0.221, 0.301]	0.187 [0.127, 0.246]	0.266 [0.204, 0.359]	0.307 [0.209, 0.404]	0.235 [0.171, 0.312]

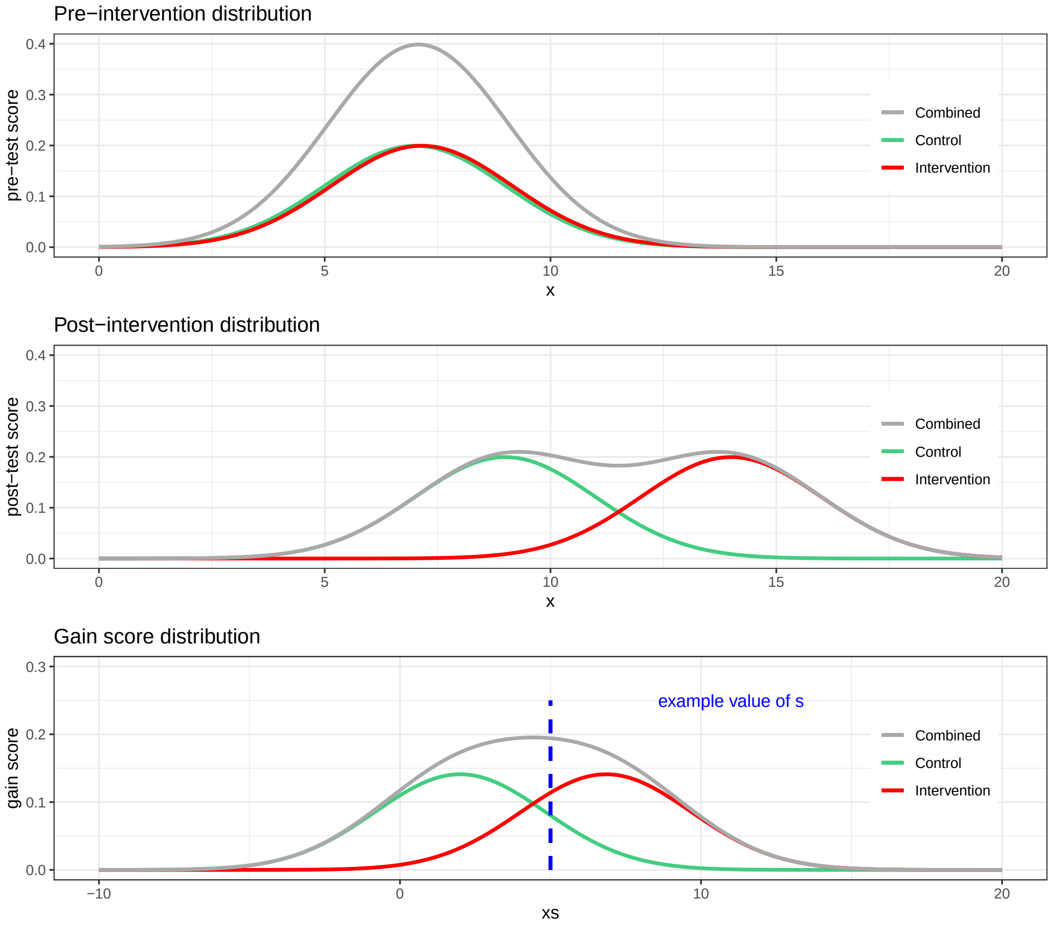

Figure 1: Pre- and Post-intervention distributions and gain score distribution. The value s is a hypothetical threshold in the notation of Section 2.2.

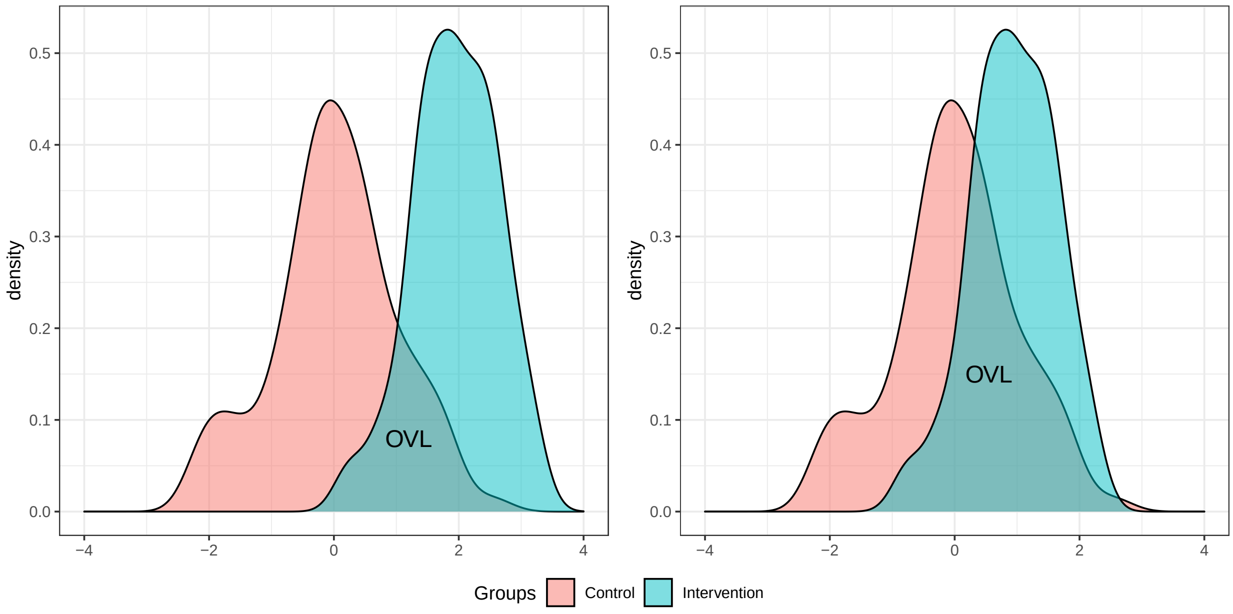

Figure 2: Low OVL (left) and high OVL (right).

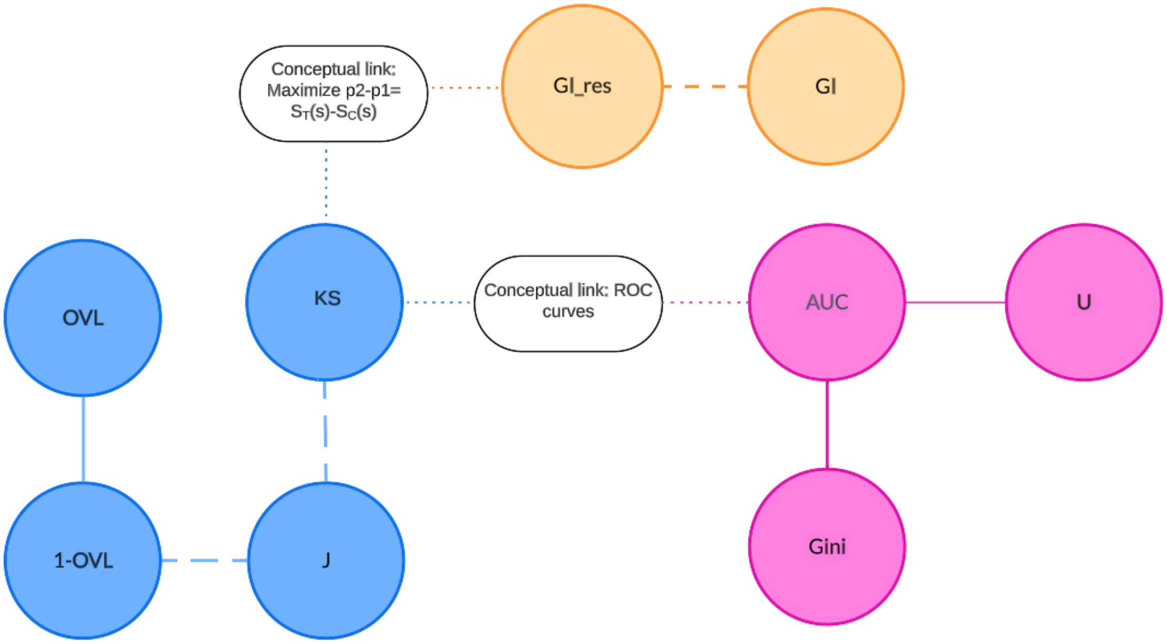

Figure 3: Relationship between several criteria to assess the agreement of two distributions. Solid lines mean equivalence up to a linear transformation; long-dashed line mean equivalence under stochastic dominance; short-dashed lines mean different handling of the multilevel (clustering) structure; dotted lines indicate conceptual relationship.

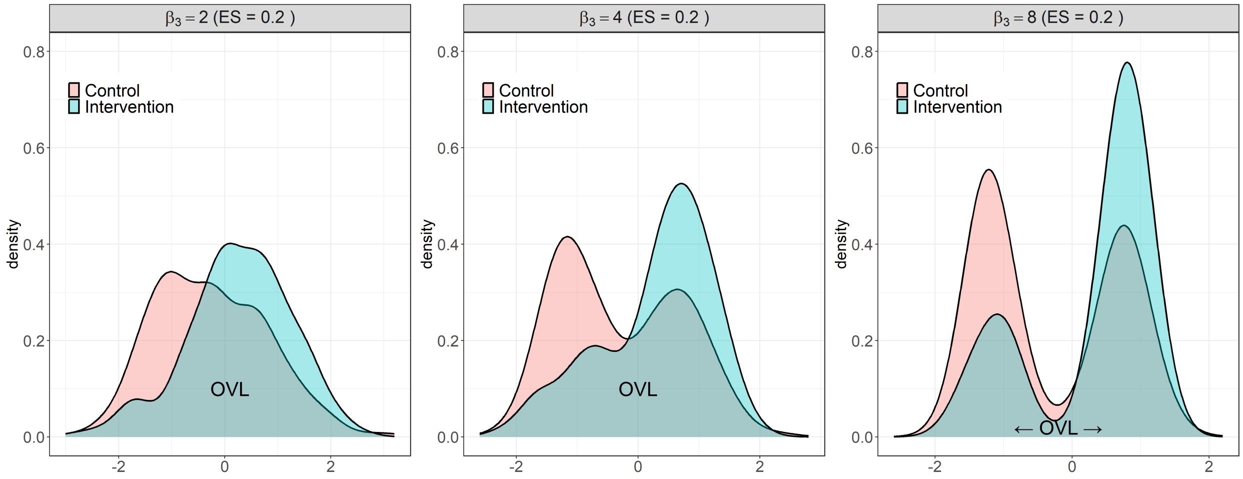

Figure 4: Overlapping figures with different values of β_3 representing how well the latent groups (LG) are separated. Each of the three panels show a (single) simulated data set of size 480 consisting of intervention and control groups with ES $= 0.2$. Other parameters are set as described Section 4.2.

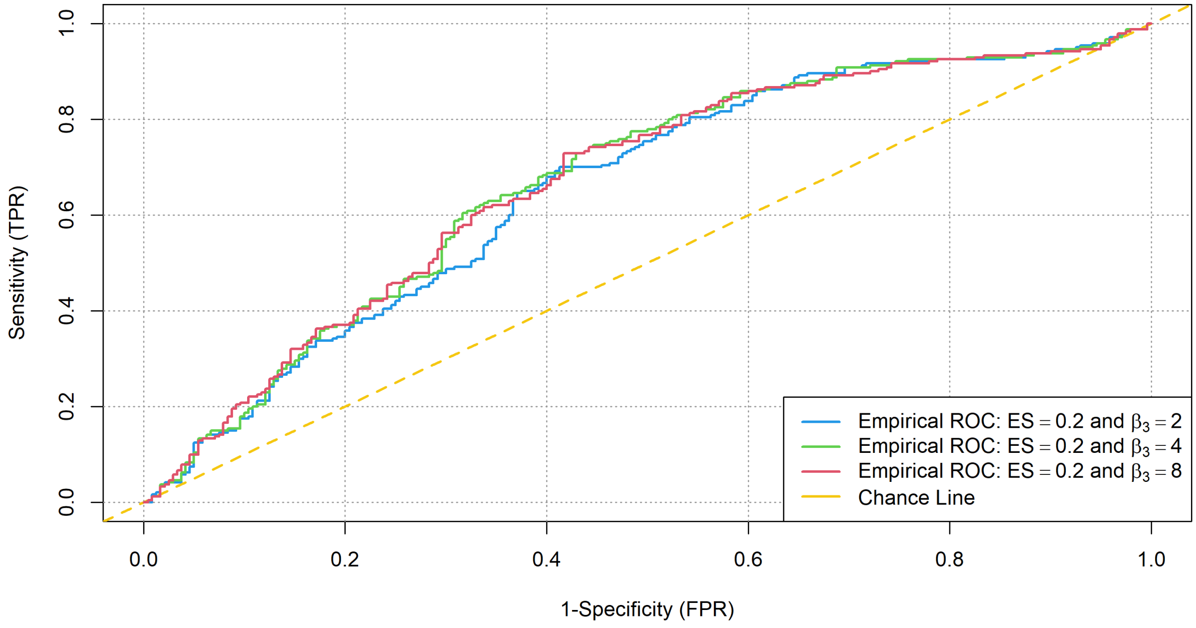

Figure 5: ROC curves for the same simulated data sets as in Figure 4 with different values of β_3, representing how well the latent groups (LG) are separated.

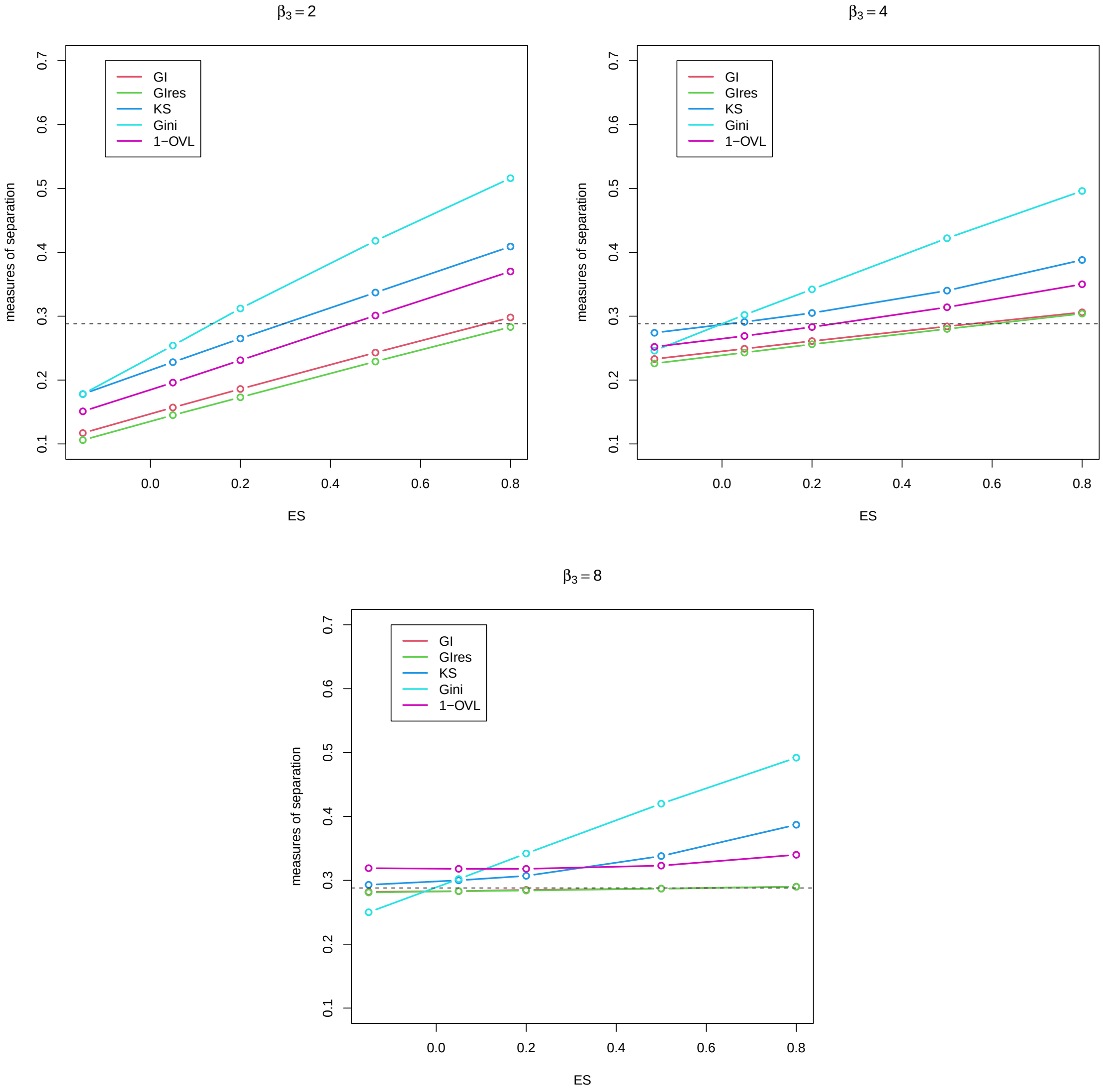

Figure 6: Top left, top right, and bottom panels: Measures of separation as a function of the true ES, for fixed β_3. The true gain index (at 0.288) is indicated by a dashed line.

List of figure captions

Figure 1: Pre- and Post-intervention distributions and gain score distribution. The value s is a hypothetical threshold in the notation of Section 2.2.

Figure 2: Low OVL (left) and high OVL (right).

Figure 3: Relationship between several criteria to assess the agreement of two distributions. Solid lines mean equivalence up to a linear transformation; long-dashed line mean equivalence under stochastic dominance; short-dashed lines mean different handling of the multilevel (clustering) structure; dotted lines indicate conceptual relationship.

Figure 4: Overlapping figures with different values of β_3 representing how well the latent groups (LG) are separated. Each of the three panels show a (single) simulated data set of size 480 consisting of intervention and control groups with ES $= 0.2$. Other parameters are set as described in Section 4.2.

Figure 5: ROC curves for the same simulated data sets as in Figure 4 with different values of β_3, representing how well the latent groups (LG) are separated.

Figure 6: Top left, top right, and bottom panels: Measures of separation as a function of the true ES, for fixed β_3. The true gain index (at 0.288) is indicated by a dashed line.