RESEARCH-ARTICLE

A Framework for Simulating Subjective Experiments: Testing Subject Screening

Accepted: 07 December 2025
Revised: 25 July 2025
Received: 31 January 2025

A Framework for Simulating Subjective Experiments: Testing Subject Screening

IRENE VIOLA*, Centrum Wiskunde en Informatica, The Netherlands
LUCJAN JANOWSKI*, AGH University of Science and Technology, Poland

The ITU-T recommendations BT.500 and P.910 outline multiple subject screening methodologies for subjective multimedia quality experiments. Yet, their real-world effectiveness remains difficult to verify due to the lack of known ground truth. This paper introduces a comprehensive simulation framework designed to objectively assess subject screening methods by generating synthetic subjective scores with known parameters. Two primary experimental scenarios — typical and super-precise subject models — were evaluated using simulated data. Results indicate that correlation-based screening methods (P.910) outperform kurtosis-based methods (BT.500) in detecting irrelevant subjects, thereby improving the precision of subjective experiment outcomes. Additional contributions include the development of a novel score generation model and the definition of robust evaluation metrics. We hope this paper will serve as the basis for future analysis based on simulations of subjective experiments.

CCS Concepts: • **Human-centered computing** → **Laboratory experiments**; **User studies**; **User models**; • **General and reference** → **Evaluation**; *Metrics*.

Additional Key Words and Phrases: Subjective experiments, simulation, subject screening, outlier detection, quality assessment

1 Introduction

In the field of multimedia research, subjective experiments are a cornerstone for evaluating the quality perceived by a user. These experiments are essential because multimedia systems and applications are ultimately designed for human consumption, and objective metrics alone often fail to capture the complexities of human perception [17]. By carefully conducting subjective experiments, researchers can collect high-quality data that reflects genuine user responses. The validity of the conclusions drawn from these studies is based on the reliability and representativeness of these data. Poorly designed experiments or low-quality data can lead to misleading results, compromising the development and optimization of multimedia systems. Thus, ensuring the collection of accurate and meaningful subjective data is critical to advancing research and improving multimedia technologies.

To ensure the quality of subjective data, an effective handling of outliers is needed. In most cases, studies are interested in an average user opinion, which is not reflected by outliers. Those outlying subjects can significantly skew results and lead to erroneous conclusions. Outliers in subjective experiments can arise from factors such as inattention of participants, misunderstanding of tasks, extreme personal biases, or technical problems [12]. Detecting and managing these outliers requires robust methodologies tailored to the specific characteristics of the data. In practice, researchers often resort to standardized methods for outlier detection. In particular, the International Telecommunication Union Telecommunication Standardization Sector (ITU-T) Recommendations

*Both authors contributed equally to this research.

Authors' Contact Information: Irene Viola, irene.viola@cwi.nl, Centrum Wiskunde en Informatica, Amsterdam, The Netherlands; Lucjan Janowski, AGH University of Science and Technology, Krakow, Poland, lucjan.janowski@agh.edu.pl.

ACM 1551-6865/2025/12-ART
https://doi.org/10.1145/3785667

BT.500 [7] and P.910 [8] detail three subjective screening methods that can be used for detecting and correcting outlier behavior in subjective data.

Testing the accuracy of the subject screening methods on real subjective data is not possible, as it is not known in advance which subjects, if any, would exhibit outlying behavior. Synthetic datasets obtained through simulations, on the other hand, would allow accurate testing of specific subject screening methods by precisely controlling the number of outlying subjects and the percentage of scores affected by outlying behavior. Therefore, incorporating simulations into the evaluation process is crucial to advance the development and validation of reliable subject screening methods. However, the methods described in the recommendations have not been rigorously tested by simulation, as models of how subjects behave in a typical subjective experiment are relatively new [10, 12], and no simulation software is available to create synthetic datasets. Thus, no benchmarking of such methods is available in the literature. This is the research gap we aim to address in this work.

In this paper, we compared the three subject screening methods described in the BT.500 and P.910 with respect to using all available data (i.e., performing no outlier detection or removal). To do so, we defined a simulation framework for subjective scores that allowed us to effectively generate realistic subjective scores, and then to manipulate the amount of randomness added by a specific subject, as well as the number of subjects that generate randomness. Doing so, we could rigorously test different methods on large amounts of synthetic data. As no large-scale benchmarking of subject screening methods has ever been performed, no metrics have been defined to evaluate their performance. To fill this gap, we propose several metrics to assess the effectiveness of the methods according to four categories: error, accuracy, dispersion, and resolution. Our results indicate which outlier detection method would be more suitable in different scenarios.

The list of contributions is as follows:

(1) We propose a simulation algorithm generating subjective scores. We envision two scenarios: 1) a typical subjective experiment scenarios; 2) a scenario with little variation among subjects ("super precise"). The two scenarios allow testing the performance of the subject screening methods in different situations, where it might be easier or harder to spot outliers. The algorithm is implemented in open-source software (available here: https://github.com/LucjanJanowski/subjective-exp-simulation-subject-screening) that allows researchers and practitioners to perform their own simulations and extend them with different subject scenarios.
(2) We propose 10 metrics to evaluate the performance of subjective screening methods, based on their capability of correctly detecting outliers, inferring the true quality value, minimizing the dispersion of the data, and aiding in statistical analysis. The metrics provide several axes on which to evaluate the performance of the subject screening methods, and can be used to benchmark future methods as well.
(3) We evaluate the methods described in the standards, showing their strengths and shortcomings. Such in-depth evaluation has never been conducted before, as the nature of subjective data does not allow to know in advance how many outliers are present in a given dataset. Thanks to our simulation framework, we are able to showcase the performance of different outlier detection and removal algorithms as more and more outliers are added to the data. In particular, we indicate how one algorithm should not be used as it fails to improve the results in the majority of cases.

2 Related work

2.1 Standardized methods for subject screening

In ITU-T recommendations, a set of standardized procedures are defined to evaluate the quality of telecommunication services, networks, and systems through rigorous and reproducible testing methods. By following the procedures detailed in the standards, researchers and practitioners can be relatively sure to minimize bias in their evaluation and obtain reliable results. As part of the standards, "post-screening" methods can be defined to detect

and possibly eliminate subjective data that is presumed not to represent the general behavior of the population and would thus skew the results.

Three main subject screening methods are standardized in ITU-T Recommendations. ITU-T Recommendation BT.500 [7] describes two of them: the first is Kurtosis-based (we will refer to it as BT.500 in the rest of the paper), whereas the second is correlation-based and refers the readers to ITU-T Recommendation P.910 (hence we will refer to it as P.910 in the remainder of the paper). ITU-T Recommendation P.910 also describes the Alternative Projection (AP) method, which is not a typical screening method since it does not detect scores to be removed but uses the weighting method, limiting influences of answers coming from less relevant subjects. In the following, each screening method is briefly described, mostly by equations. For more details, please link to specific Recommendations.

In order to describe screening methods, we start by describing single-subject answer. In BT.500, a subjective score is described as u_{ijkr} for i-th subject, j-th test condition (PVS), k-th sequence (SRC), and r-th repetition. For simplicity, in this paper, we assume that all subjects rated all PVSs and we ignore sources, as they are not used in the screening methods. Thus, a score is given by u_{ij} with I users and J PVSs.

2.1.1 Kurtosis-based screening: BT.500. Method BT.500 is based on counting the number of times a specific answer is out of range. Specifically, two counters P_i and Q_i are computed per subject i, by iterating over every PVS j:

$$\text{if} \quad u_{ij} \geq \bar{u}_j + k(\beta_j^{(2)}) \cdot S_j \quad P_i = P_i + 1, \tag{1}$$

$$\text{if} \quad u_{ij} \geq \bar{u}_j - k(\beta_j^{(2)}) \cdot S_j \quad Q_i = Q_i + 1, \tag{2}$$

in which $\bar{u}_j$ is the average score for PVS j, and $S_j = \sqrt{\sum_{i=1}^{I} \frac{(u_{ij}-\bar{u}_j)^2}{I-1}}$ is the standard deviation.

The coefficient k depends on the kurtosis coefficient $\beta_j^{(2)} = \frac{m_4}{(m_2)^2}$ where m_x is given by:

$$m_x = \frac{\sum_{i=1}^{I}(u_{ij} - \bar{u}_j)^x}{I} \tag{3}$$

The kurtosis coefficient indicates the tailedness of the score distribution; thus, it can be used to ascertain whether the scores can be considered normally distributed or not. The kurtosis value is here used to determine the value of coefficient k as such:

$$k(\beta_j^{(2)}) = \begin{cases} 2, & \text{if } 2 \leq \beta_j^{(2)} \leq 4 \\ \sqrt{20}, & \text{otherwise} \end{cases}$$

Finally, after calculating P_i and Q_i for all scores given by a specific subject, we test two conditions: $\frac{P_i+Q_i}{J} > 0.05$ and $\left|\frac{P_i-Q_i}{P_i+Q_i}\right| < 0.3$. If both are true, then we reject subject i.

2.1.2 Correlation-based screening: P.910. The method we call P.910 is based on Pearson's correlation between subjective scores and mean for all scores:

$$\rho_i = cor(u_{ij}, \bar{u}_j) \tag{4}$$

A threshold r is set for the correlation (in the standard, $r = 0.75$). If $\rho_i < r$, then the subject is discarded, and the new average $\bar{u}_j$ is computed by excluding the subject. The process is repeated until no ρ_i is below the threshold.

2.1.3 Alternating projection screening. Recommendation P.910 describes another method based on subject model "Bias-subtracted consistency-weighted MOS method for subject screening". It employs a recursive AP process using calculated parameters to improve the precision of calculating parameters in the next step. Three model parameters are calculated by:

$$\mu_j^{(l)} = \frac{\sum_{i=1}^{I}((\sigma_i^{(l-1)})^{-2}(u_{ij} - \mu_{\Delta_i}^{(l-1)}))}{\sum_{i=1}^{I}(\sigma_i^{(l-1)})^{-2}} \tag{5}$$

where $\sigma_i^{(0)} = 1$ and $\mu_{\Delta_i}^{(0)} = 0$

$$\mu_{\Delta_i}^{(l)} = \frac{\sum_{j=1}^{J}(u_{ij} - \mu_j^{(l)})}{J} \tag{6}$$

$$r_{ij} = u_{ij} - \mu_j^{(l)} - \mu_{\Delta_i}^{(l)} \tag{7}$$

$$\sigma_i^{(l)} = \sqrt{\frac{\sum_{j=1}^{J}\left(r_{ij} - \bar{r}_i\right)^2}{J-1}} \tag{8}$$

The process is run until convergence is achieved (i.e., $\sum_{j=1}^{J}(\mu_j^{(l-1)} - \mu_j^{(l)})^2 < \epsilon$). Afterwards, $\bar{u}_j \leftarrow \mu_j^{(l)}$. Please note that the method does not modify the distribution of the scores, but updates the mean to give less weight to untrustworthy subjects. Thus, it cannot be directly used for statistical tests that require the distribution of scores, such as ANOVA or t-test.

2.2 Non-standardized methods

Besides methods described in the standards, not much work is related to subject screening. In [11] adaptation of Generic Probabilistic Model (GPM) for subjective data was proposed. This approach is similar in the concept to AP method. The main difference is using a different underlying model and assumption that a subject change states from spamming to correct answers. Since in this work we are interested in testing solutions described in standards, GPM method was not implemented. We leave it as a future research.

An interesting problem related to subject screening in psychology is described in [19]. It is pointing that by removing study participants, we introduce specific bias. Again, understanding better how a specific screening method works is crucial for better understanding eventual bias introduction. This work was done for the measurement of response time in [1] where the simulation study allows comparing three different methods used in the psychology study. The goal of our work is a similar analysis in the multimedia quality domain.

The main goal of screening subjects is to obtain more precise data. The problem of multimedia data precision is described in numerous publications. One of the first attempts in this direction was the proposal of the SOS coefficient [6]. The extension of this concept using subject models and simulations is described in [9]. The most systematic analysis of subjective experiments is presented in [17]. This publication extends the ideas presented in [9].

3 Simulation methodology

The recent development of the subjective model allows the simulation[1] of a typical subjective experiment [10, 12, 15]. In this paper, we use a model proposed in [12], because it contains parameters that describe a subject

[1] generate scores for I subjects and J stimuli

and a stimulus and provides a fast estimation algorithm. For simplicity, we refer to this model as Zhi2020. For the Zhi2020 model, the subjects' scores are given by:

$$u_{ij} = \min(5, \max(1, \lfloor \psi_j + \Delta_i + \mathcal{N}(0, \sigma_i) \rceil)) \tag{9}$$

where ψ_j is the true quality of a stimulus j, σ_i is the precision of the ith subject, Δ_i is the bias of the ith subject, and $\lfloor x \rceil$ is integer the closest to x.

Consequently, when simulating an experiment with I subjects and J stimuli, one has to select $J + 2I$ parameters. However, the selection method for these parameters is not described in the literature. Here, we propose a parameter selection procedure to simulate a typical laboratory experiment and the so-called "Super Precise" experiment. We start from selecting $2I$ subjects' parameters.

3.1 Subjects' Parameters

The behavior exhibited by subjects participating in user studies naturally spans a wide range of behaviours. For simplicity, in our work we describe two types of subjects: a "Typical" subject and a"Super Precise" subject. We argue that on top of the "Typical" subject, the "Super precise" subject simulation is especially useful in the scope of analyzing subject screening methods.

The "Typical" subject simulates a subject observed in a subjective experiment. Therefore, we reanalyzed all subjective experiments described in [14] using the AP estimation algorithm described in Section 2.1 [12]. We excluded experiments with repeated scores. We obtained 423 estimates of the bias and sigma parameters. Next, we use the R package `fitdistrplus`[2] to find the parameters to fit the 423 observed samples to a normal distribution. Doing so, we obtained:

$$\Delta_i \sim \mathcal{N}(0, 0.3375)$$

and

$$\sigma_i \sim \text{Lognormal}(-0.431, 0.191)$$

The parameters of the "Typical" subject were estimated based on data from subjective experiments, which can include scores of subjects that should be screened. Therefore, we extended our simulations by including "Super Precise" subjects. Intuitively, "Super Precise" subject has no bias and no variance of the ratings, i.e. $\Delta_i = 0$ and $\sigma_i = 0$. However, in such a case, we cannot obtain a MOS different from a discrete value. This is because equation (9) for $\Delta_i = 0$ and $\sigma_i = 0$ is:

$$u_{ij} = \lfloor \psi_j \rceil$$

which is a discrete value[3].

To solve this problem, we have found σ_i such that:

$$\min_{\sigma} \left(E(\lfloor \mathcal{N}(\psi, \sigma) \rceil) - \psi \right)^2$$

The minimum for the above equation, obtained by simulations, is reached for $\sigma = 0.36$. Finally, "Super Precise" subjects are defined by:

$$\Delta_i \sim \mathcal{N}(0, 0.01)$$

and

$$\sigma_i \sim \text{Lognormal}(\ln(0.36), 0.01).$$

As mentioned before, many other subject models can be crafted, either by theoretical modelling or by obtaining subject parameters from other existing datasets. Our framework allows adding to the existing models to test the outlier detection and removal algorithms in more diverse settings.

[2]https://cran.r-project.org/web/packages/fitdistrplus/index.html

[3]More details on the behavior of Zhi2020 for different variances can be found in [2].

3.2 Stimuli Parameters

In Zhi2020 model, a stimulus is described by one parameter ψ representing the true qualities. A well-designed experiment should be designed so that the true qualities cover the entire range of quality scores from 1 to 5. The easiest solution would be to draw the true qualities from a uniform distribution $\psi_j \sim U(1,5)$. However, a typical set of experimental stimuli is generated by selecting source videos and compressing them at different levels. These levels are due to different compression settings and/or different codecs used for compression. Therefore, we have to select the quality of the source stimuli and determine how the quality drops with compression.

In an ideal scenario, the ψ distribution of the source stimuli would be equal to the highest value of quality score, i.e., 5, since the source stimuli are uncompressed and undistorted. However, in practice, the source stimuli can have varying scores, due to subjective preference or acquisition errors in the source itself [16]. Nonetheless, the distribution ψ of the source data will be skewed towards the higher ranges of the scale, and low values will be given very seldom. To accurately simulate real experiments, we derive the ψ distribution of the source stimuli from the MOS obtained for the quality of the sources in existing subjective experiments. We used the HDTV test [16] to estimate the quality of the 78 sources. We use the R package `fitdistrplus` to fit the 78 sources to a beta distribution. The beta distribution was chosen as it allows for the modeling of random variables over a finite interval (in our case, the minimum and maximum value of our test, e.g., $[1,5]$). Doing so, we obtained:

$$\psi_k \sim 4\mathcal{B}e(20.8, 2.6) + 1$$

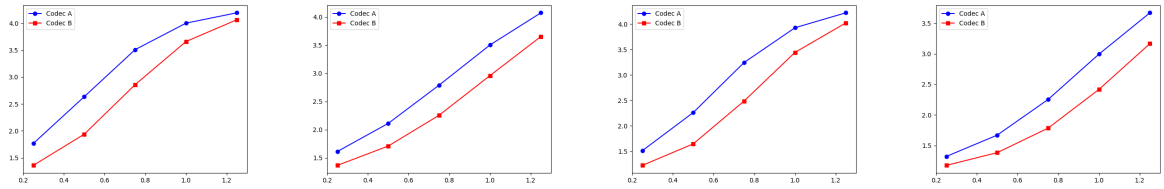
Note that the above equation provides only the source quality, i.e., the quality of the undistorted, uncompressed stimulus. In order to generate the true qualities for the whole experiment, we have to specify how the source qualities change at different compression levels. The change should follow a logistic function [5, 18]. Different source stimuli result in different shapes of the quality drops, but within a reasonable range. In addition, we need to simulate the effects of different codecs. To fulfill the above-mentioned conditions, we use the function given by:

$$\psi_j = (\psi_k - 1)\frac{e^{a_k(x-b_k+c_l)}}{1+e^{a_k(x-b_k+c_l)}} + 1 \tag{10}$$

where ψ_j is the jth stimulus quality obtained for source k, ψ_k is the true quality of the source k, x is a bit rate, a_k is a parameter that determines the slope of the logistic function, b_k is a parameter that determines the position of the logistic function, c_l is a difference between codecs.

To obtain different sequences and codec behaviors, the parameters of equation (10) are random. We selected, by simulations, such distributions $a \sim U(3,6)$, $b \sim U(0.3,1.2)$. The only parameter that is not random is c, which models the shift between the sigmoid functions generated for codecs 1 and 2. Our objective is to determine the difference in quality for these codecs. This difference changes with x. Our simulations show that the difference in quality is linearly related to c if $c < 0.5$. A higher c can generate an unrealistic ψ_j distribution. Based on our simulations, given that $c < 0.5$, the mean difference is: mean_dif $= 2.6c$ and the maximum difference is max_dif $= 4.5c$.

Let us analyze an example to clarify the algorithm described above. Consider a simulation with 16 source stimuli, each compressed with two codecs at five levels of compression. The true quality for codec A is higher than for codec B on average by 0.5. The simulation involves four steps. The first is to draw 16 values from $4\mathcal{B}e(20.8, 2.6) + 1$ - let us call them $\psi_1, \cdots, \psi_{16}$. These are the source quality values, when no distortion is applied. The second step is to draw parameters for specific source stimuli $a_1, \cdots, a_{16}$ and $b_1, \cdots, b_{16}$ from specific distributions. The third step is to calculate the true qualities for codec A for the selected parameters $\psi_k, a_k, b_k, c_1 = 0$ and $x = \{0.25, 0.5, 0.75, 1.0, 1.25\}$ (the values of x are chosen to mimic equally-spaced bitrate values). The final step is to calculate the true qualities for sequences using codec B. Since the mean quality difference between codecs A and B is 0.5, $c = \frac{0.5}{2.6} = 0.192$. Therefore, the qualities of codec B are calculated with the parameters



(a) (b) (c) (d)

Fig. 1. Example of true quality curves for a two-codec scenario, using the parameters detailed in Sec. 3.2. The difference between the quality of the two codecs is set to 0.5.

$\psi_k, a_k, b_k, c_2 = 0.192$ and the same x. An example of four random curves obtained with these parameters can be seen in Figure 1.

3.3 Validation

To validate our stimuli generation framework, we use the HDTV dataset 3 [16]. In particular, we extract the MOS scores for each PVS, and we estimate the subject parameters using the alternating projection algorithm [12]. Then, we use our framework to generate the corresponding subjective scores. We run 1000 simulations and we compute the Pearson correlation coefficient ρ and the RMSE between the original MOS and the MOS obtained from our simulation. Results show high consistency with the original scores, with an average $\rho = 0.992$ (std = 0.00083), and an average RMSE = 0.147 (std = 0.00761).

3.4 Experimental scenarios

In the simulation, we assume that the goal of the experiment is to compare two different codecs. We consider two scenarios: Scenario A: Typical subjects and Scenario B: Super Precise subjects. Both scenarios refer to typical lab experiments with 160 stimuli (16 sources, two different codecs, each with five different levels) and 24 subjects. In both cases, ψ is generated as described in Section 3.2. In cases where simulation considers different qualities, we specify the selected value of c. For other simulations, we select $c = 0$[4]. The difference between scenarios is the subject's precision. Scenario A and Scenario B differ in terms of subject parameters. Those parameters are generated from typical subject distributions or Super Precise subject distribution, respectively, as described in Section 3.1.

To evaluate how subject screening methods work, we need to add outliers to our scores. In this manuscript, we decided to limit our analysis to a simple scenario in which erroneous answers u_{ij} are assigned to a certain percentage p of the set of PVSs J by a certain number N of the total set of subjects I, due to incorrect labeling of the results, as seen in [12]. To do so, we permute the scores of N subjects with probability p.

4 Metrics

Defining what constitutes an outlier is a nearly impossible task [19]. However, at the bare minimum, we want our outlier detection algorithm to be able to identify erroneous answers, which might be due to an incorrect labeling of the results, as shown in [12]. Thanks to our simulation setup, we have the ability of knowing exactly what answers were erroneous. Thus, we define two *error* metrics: the True Detection Probability (TDP), which indicates the probability of a subject screening method to actually detect 100% random answers, and the False

[4]No difference between codecs.

Detection Probability (FDP), which indicates the probability of a subject screening method to flag admissible answers as outliers.

Each outlier detection algorithm can potentially change the estimated *accuracy* of the MOS and the *dispersion* of the obtained value. Both the *accuracy* (i.e., how far the estimated value is from the true value) and the *dispersion* (i.e., how the obtained values are statistically distributed) can influence the experiment *resolution*, that is, the ability to see statistical differences for specific real differences. Therefore, to compare different outlier detection algorithms, we have to use metrics that target all of those aspects.

To check the accuracy of our estimators from the true quality when outliers are added (or removed), we use three methods. Firstly, we want to check the linear relationship between our estimated value MOS_γ (where $\gamma = [\text{None, BT.500, P.910, AP}]$ denotes the method chosen to estimate the MOS value), and the true quality ψ. To do so, we use the Pearson Linear Correlation Coefficient (PLCC). Similarly, we check the monotonicity of the relationship between ψ and MOS_γ using Spearman's Rank Order Correlation Coefficient (SROCC). Finally, to have an idea of the error between our estimation and the true value, we compute the Root Mean Square Error (RMSE). Higher values of PLCC and SROCC, and lower values of RMSE, will denote more similarity between the true value and its estimation.

Subject screening algorithms can potentially change the distribution of the scores from which the estimated value is drawn, thus changing the results' distribution. In order to quantify it, we rely on three metrics. The first is the Standard Error (SE) estimation, which is commonly used in order to compute the Confidence Intervals (CIs). The SE allows us to measure the dispersion of our estimation, weighted by the size of the sample. A smaller value of SE indicates narrower CIs; however, that alone does not guarantee a more precise estimation, as the true quality value might lie outside of the SE range. Thus, we measure whether the true quality value would lie in the CIs computed based on our estimation algorithms, and we denote it with Confidence Interval Accuracy (CIA). This gives us an indication of whether our estimation algorithms significantly changed the distribution, to the point that the true quality would not be captured. Drawing from the literature, we also compute the SOS a parameter [6], to understand whether adding (or removing) outliers causes notable changes in its value. The SOS a gives an idea of similarity of the users, and the reliability of the user study.

Subjective studies are often used to judge whether one source of distortion is to be preferred with respect to another: for example, to select the compression configuration that leads to the highest MOS. To perform the judgement, Student's t-tests are commonly employed, with the null hypothesis stating that the two configurations are equivalent, and the alternative hypothesis stating that they are different. In Student's t-tests, two type of errors can be found: type I (false positive) errors occur when the null hypothesis is wrongly rejected; type II (false negative) errors occur when we fail to reject the null hypothesis, even though the hypothesis is false. The probability of committing a type I error is denoted with α, whereas the probability of a type II error is denoted with β: the first is set in advance by the investigator, whereas the second depends on the size of the samples and their differences. Thanks to our simulation, we are able to know in advance whether the null hypothesis should be rejected or not; this gives us the opportunity to test the effect of the subject screening methods on the α and β errors[5]. To do so, we perform a t-test, and we count the number of times the hypothesis is correctly rejected or not; this further gives us an idea of the resolution of our estimations.

We list all the metrics, with relative marker, in Table 1.

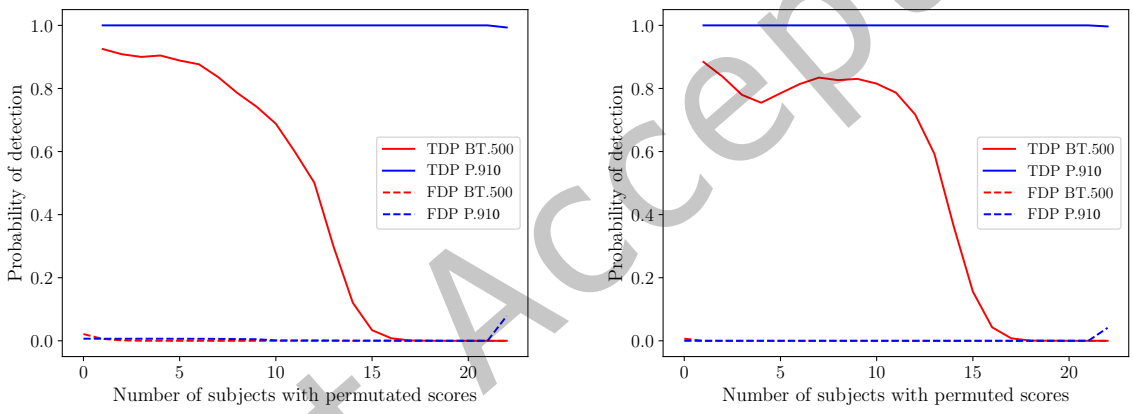
5 Results

The results are divided by the category, as described in Table 1.

[5]Note that, since they refer to different scenarios, at any given point we will have *either* the α error (if we set the configurations to be equivalent) or the β error (if we set the configurations to be different)

Table 1. Metrics used to compare different subject screening methods.

Metric	Category	Description
TDP	error	Probability of detecting 100% random answers
FDP	error	Probability of detecting "normal" tester as a random
PLCC	accuracy	Testing linearity between MOS and ψ
SRCC	accuracy	Testing ordering between MOS and ψ
RMSE	accuracy	Testing difference between MOS and ψ
SE	dispersion	Average standard error of each PVS for the whole experiment
CIA	dispersion	Probability that ψ is in the 95% confidence interval
SOS-a	dispersion	Overall precision of the experiment
α	resolution	Significance level: Probability of wrongly rejecting the null hypothesis
$1-\beta$	resolution	Power: Probability of correctly rejecting the null hypothesis



(a) Scenario A: Typical subjects

(b) Scenario B: Super Precise subjects

Fig. 2. True and False Detection Rate for outlier detection algorithms BT.500 and P.910, as a function of the number of fully random subject scores. AP and None methods do not mark subjects to be removed, therefore TDP and FDP cannot be calculated.

5.1 Error

Results of the simulations with scenario A and B are presented in Figure 2. Specifically, we show the TDP and FDP (see Table 1) for outlier detection algorithms BT.500 and P.910, for different number N of subjects whose scores were permuted. In this simulation scenario, we permute the N subjects with probability $p = 1$. Please note that we could not use the error metric on the AP algorithm, as the algorithm does not flag outliers.

An ideal outlier detection algorithm should be able to always identify fully randomized answers from a subject (TDP = 1), while avoiding to flag "correct" answers as outliers (FDP = 0). For Scenario B (*Precise* subject model), we can see that BT.500 does not work, as it is unable to reach a perfect detection of subjects with permuted scores: even when only one subject has been affected with score permutation, the probability of detection remains at

around 0.9. We notice a peculiar dip in TDP value when $N = 4$, before increasing as N increases up to $N = 10$. The dip can be explained by considering the nature of the score distribution and how the BT.500 algorithm works (see equation (3)). As our subject model has very low variance, when little or no randomization is applied to the scores, it is unlikely that, for a given stimulus, the score distribution will be considered normal (that is, with a kurtosis coefficient between 2 and 4). As N increases, the score distribution for a given stimulus starts converging to a normal distribution, thus leading to a change in coefficient. This explains why, from $N \geq 5$, we observe an increase in TDP values. However, from $N = 9$ onward, the TDP sharply decreases until $N = 17$, where we observe that BT.500 is unable to detect any of the outliers. This can be explained considering again the nature of the algorithm. As the number of users with permuted scores increases, the distribution of the scores for a given stimulus approximates a uniform distribution between 1 and 5; thus, the average score $\mu \rightarrow 3$, and the standard deviation $\sigma \rightarrow 2/sqrt(3)$. It is statistically unlikely that a user with randomized scores will fall far outside of the distribution often enough to be flagged as an outlier. On the other hand, we can notice how P.910 is more robust to such randomization: all the outliers are correctly flagged as such, even when they represent the majority of the subjects. In terms of FDP, we can see that it is close to 0 for both algorithms, meaning that neither of them incorrectly flags normal subjects as outliers.

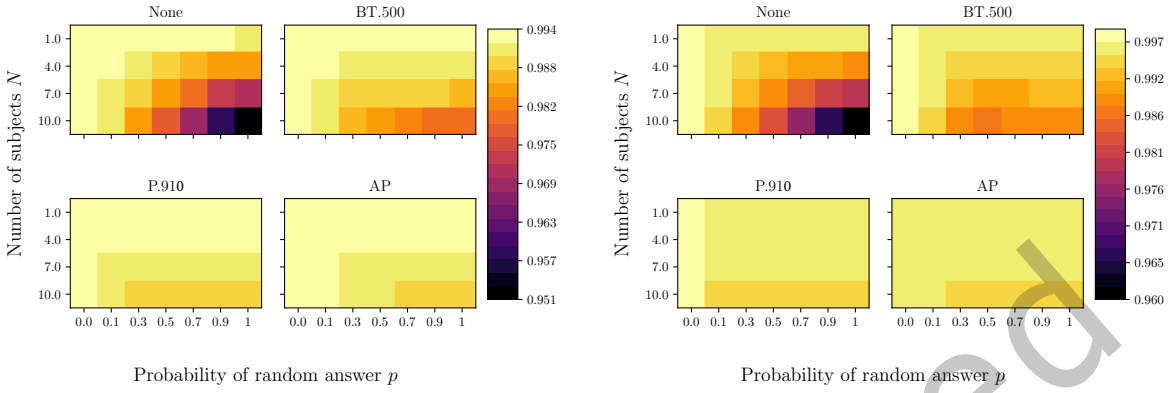
Similar trends can be observed when we consider Scenario A (*Classic* subject model). In particular, we can notice that again, BT.500 is unable to correctly flag as outliers users with permuted scores with 100% accuracy, whereas P.910 is always able to detect users with permuted scores.

5.2 Accuracy

In this section, we test the correlation between our true score ψ and the MOS we compute after applying a given outlier detection algorithm. In this case, we test all the standardized subject screening methods, and we compare it with a baseline of no subject screening. To obtain a better granularity in the percentage of scores that are affected by permutation, we now consider two parameters: the number of subjects N whose scores are affected by permutation, and the probability p of each score being permuted, as indicated in section 3.4. Please note that the $p = 0$ scenario indicates the ground-truth values of our simulation (i.e., no permutation was added). For the sake of keeping our results legible, we limit $N = [1, 4, 7, 10]$. Results for both scenarios can be seen in Fig. 3 and Fig. 4. We omitted the results for SRCC, as they were exhibiting similar trends with respect to PLCC.

One main observation that we can draw from the results in Fig. 3 is that the PLCC between the obtained MOS value and the true value ψ is very high in all cases under consideration; even when we select $N = 10$ and $p = 1.0$ (meaning 10 subjects with 100% probability of having their scores permuted), the worst PLCC value that we obtain is 0.951 for Scenario B (*Precise*) in the case of no subject screening. We obtain the worst results when no subject is removed; in particular, we see that the PLCC remains constant when only one subject is affected by the permutation, and slightly lowers for $p = 1$. As N and p increase, we see that the PLCC lowers accordingly. BT.500 follows similar trends with respect to the baseline case, albeit with less severe decrease in PLCC values; specifically, we see that the values progressively lower as we increase N and p. On the other hand, both P.910 and AP are more robust to variations of N and p: in particular, we can notice that for both algorithms, no change in PLCC values is observed up to $N = 4$ for Scenario A (*Classic*), and up to $N = 7$ for Scenario B (*Precise*). In Scenario A, we can see that for $N = 7$, the PLCC value slightly drops, achieving a similar performance with respect to the baseline case with $p = 1$, despite having a significantly larger amount of subjects affected; even with the highest values of N and p, the PLCC difference remains very small. In Scenario B (*Precise*), we can observe that a drop in PLCC values is only registered for the highest value of N. Comparing the two methods, AP is slightly more robust, as changes in PLCC values are visible for higher values of p with respect to P.910.

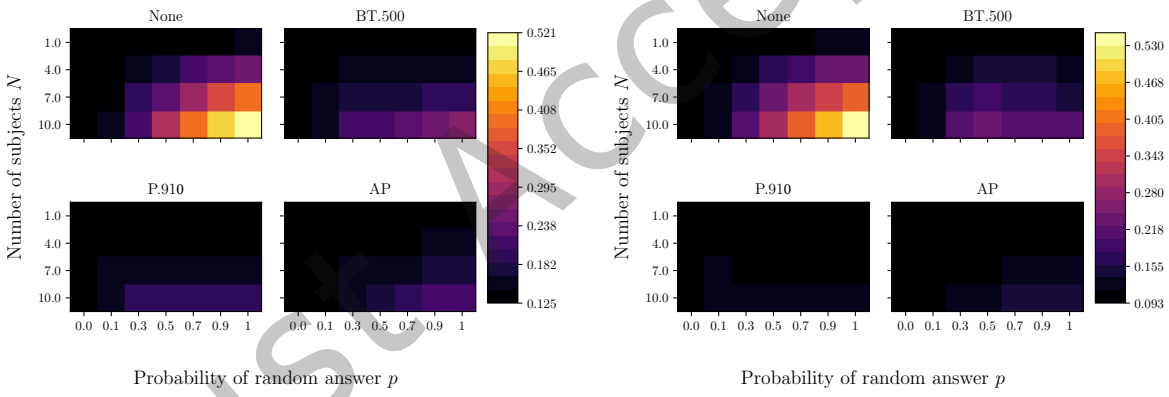
In the case of RMSE (see Fig. 4), we can immediately see the impact of not applying any screening method, as the error in the baseline progressively increases as N and p increase. In the case of $N = 10$ and $p = 1$ (meaning,



(a) Scenario A: Typical subject model

(b) Scenario B: Super-precise subject model

Fig. 3. PLCC between quality used in the simulation ψ and the observed as simulation output (MOS)



(a) Scenario A: Typical subject model

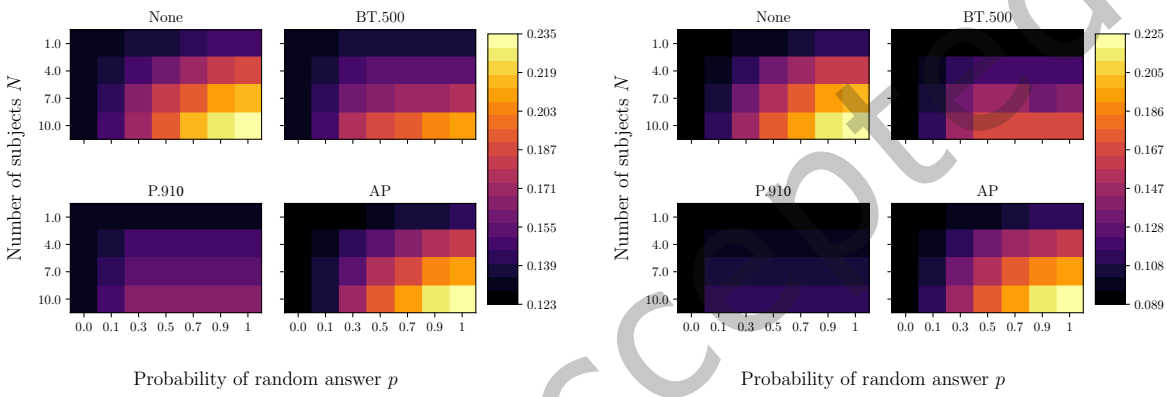
(b) Scenario B: Super-precise subject model

Fig. 4. RMSE between quality used in the simulation ψ and the observed as simulation output (MOS)

10 subjects had their scores permuted with 100% probability), we obtain an RMSE value of 0.521 in Scenario A (*Classic*), and 0.530 in Scenario B. All the subject screening methods under consideration are able to reduce the error considerably; among the three, BT.500 is the worst-performing one, with a maximum RMSE value of 0.266 in Scenario A, and 0.218 in Scenario B. AP and P.910 have very similar performance and show high robustness; the performance only degrades for $N = 10$ in Scenario A, and for Scenario B, almost no difference in RMSE is observed even for the highest values of N and p.

5.3 Dispersion

In order to test the impact of randomization on dispersion, we first check how the SE changes as we increase both the probability of a random answer p, and the number of subjects N affected by it. Fig. 5 shows the results for all the subject screening methods, for both scenarios (typical and super-precise subject model) under test. For both scenarios, we can observe an increase in SE values as N and p increase. The trend is especially clear when no outlier detection algorithm is applied: in this case, the SE almost doubles in scenario A, and almost triples in scenario B. The effect is mitigated when we apply the subject screening method BT.500 or P.910, the latter achieving a bigger reduction in SE increase. We can also observe that using the AP algorithm does not reduce the SE values; in fact, similar trends can be observed with respect to the baseline case, when no subject is removed.



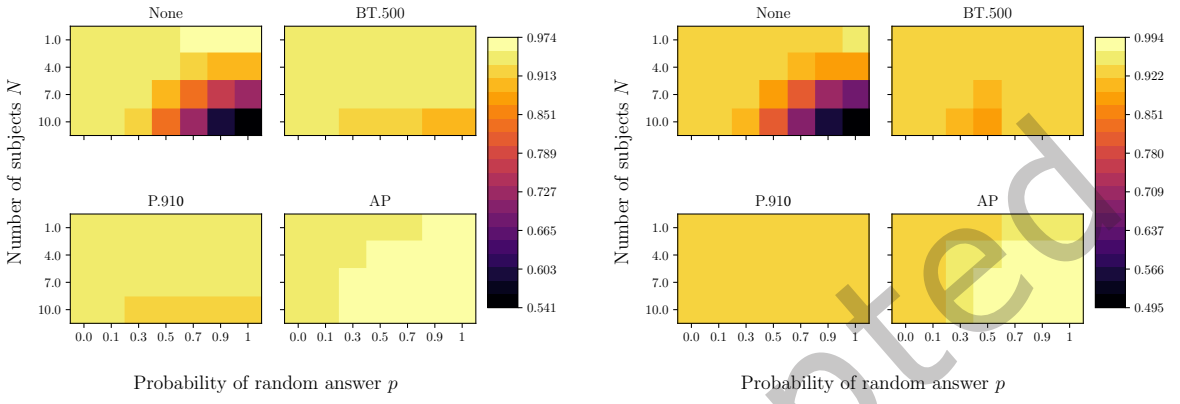
(a) Scenario A: Typical subject model

(b) Scenario B: Super-precise subject model

Fig. 5. SE between quality used in the simulation ψ and the observed as simulation output (MOS)

Fig. 6 depicts the CIA, which is the probability that the true quality ψ is within the confidence interval of the predicted MOS. This measure helps us understand whether our randomization affects the probability of statistically "capturing" the true quality, and whether outlier removal helps in that regard. In computing this interval, we are excluding the cases for which the standard error would be zero: in that case, the confidence interval would also be zero, and so would be the probability of ψ being contained in the confidence interval. We can observe that, in the absence of any randomization, the CIA is quite high, with values of 0.944 for the typical subject model (scenario A), and 0.922 for the super-precise model (scenario B). The lower values observed in the latter case can be explained by the fact that the standard error is lower in scenario B (see Figure 5), thus, the confidence intervals are narrower. We can observe that, by increasing the randomization in our scores and without applying any outlier removal or correction, the CIA gets progressively lower, despite the confidence intervals themselves getting larger, as seen by the larger values of SE. This is due to the fact that the randomization process changes the distribution of the MOS values: as we add more random noise uniformly distributed between 1 and 5, the MOS values slowly converge towards the middle of the curve. Thus, the true values ψ that are at the edges of the distribution (think of very low quality or very high quality) will naturally fall outside of the confidence interval. Applying outlier removal algorithms, such as BT.500 and P.910, reduces the SE, as seen in Fig. 5, and thus will reduce the impact on the CIA. However, despite the fact that large SE were also observed when applying AP, we do not observe the same phenomenon when looking at the CIA: in fact, in this case, the

probability of the true quality ψ being inside the confidence interval *increases*. The AP algorithm does incorporate more uncertainty in the prediction, as shown in the SE results; however, its estimation of the true quality is very robust even with high values of randomization.



(a) Scenario A: Typical subjects

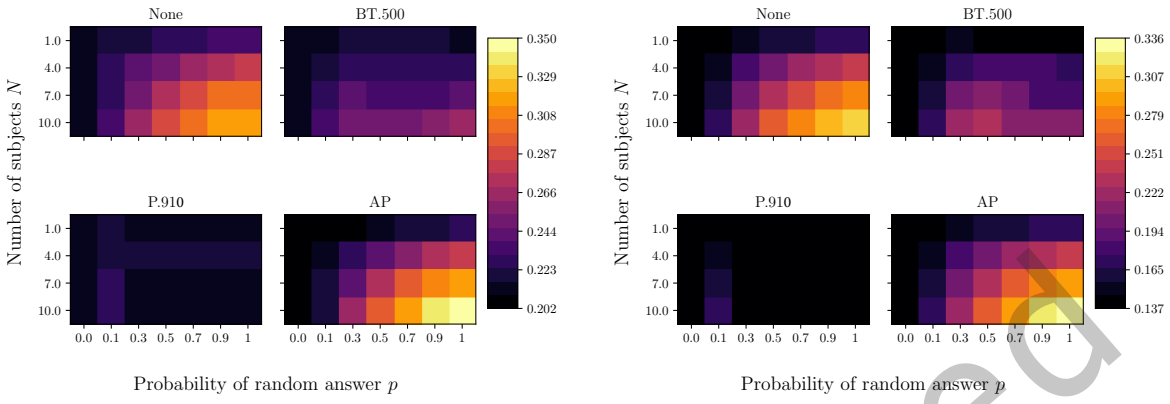
(b) Scenario B: Super precise subjects

Fig. 6. CIA between quality used in the simulation ψ and the observed as simulation output (MOS)

To see how the user distribution varies with randomization, we compute the SOS a, shown in Fig. 7. From the literature, lower values of a correspond to more precise user studies (think of controlled lab tests), whereas larger values of a are observed in less controlled conditions, such as crowdsourcing. Similar trends can be observed as we stated before: in our baseline case, as N and p increase, larger values of a are observed, in both scenario A and B. Outlier removal algorithms reduce this effect by removing the source of randomization: we can see that in both scenarios, for BT.500 and P.910, the a value remains consistent. We can also observe that this is not the case for the AP algorithm: this is to be expected, since the algorithm incorporates the uncertainty in its prediction. Thus, the a value reflects the added uncertainty and the larger differences between user models.

5.4 Resolution

In terms of resolution, we want to analyse the impact of outliers when we are using Student's t-test to compare two different sources of distortion. In this case, we need to differentiate between two cases: when the two sources are equivalent, and when they lead to different distributions in distortion. In the first case, we are interested in the *significance level* α: that is, the probability of erroneously detecting differences among the two sources when there are none. In the second case, we are interested in the *power* $1-\beta$: that is, the probability of correctly detecting differences among the two sources. In the latter case, results naturally depend on the actual difference among the sources; in other words, on the *effect size* of the test we are conducting. The effect size is linked to the power of the experiment, and its sample size: *sensitivity* analysis allows, for example, to determine the required effect size d given a certain α, $1-\beta$, and sample size n; *post-hoc* analysis allows to determine the achieved power $1-\beta$, considering the values of d, α, and n. For our experiment, in which we are comparing two distortion sources with 80 stimuli each, and 24 participants, our sample size is $n = 1920$. We consider two possible effect sizes: $d = 0.064$, which would lead to an expected power $1-\beta = 0.80$, and $d = 0.098$, which would lead to an expected power $1-\beta = 0.99$, setting $\alpha = 0.05$. All the effect sizes and relative power were computed using G*power [3, 4]. To



(a) Scenario A: Typical subject model

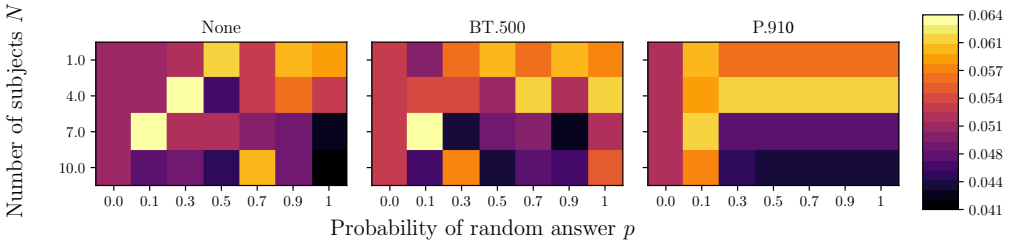
(b) Scenario B: Super-precise subject model

Fig. 7. SOS a between quality used in the simulation ψ and the observed as simulation output (MOS)

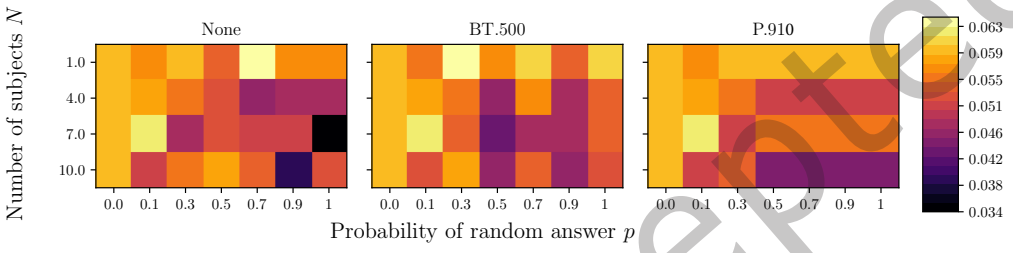
convert our effect sizes to actual differences among the curves, we consider that $d = \frac{\mu}{\sigma}$, in which μ is the mean of differences, and σ is the standard deviation of differences, which we consider coming from a continuous uniform distribution $\mathcal{U}_{[1,5]}$: $\sigma = \sqrt{\frac{1}{12}(5-1)^2}$. Please note that, since we need access to the underlying distribution of scores for each subject, we cannot test the performance of the AP algorithm, as the algorithm does not apply any change to the raw scores.

Fig. 8 depicts the variation in observed α values for different outlier detection algorithms, as a function of the probability of random answer, and the number of subjects affected by the randomization, for the two scenarios under test. We can observe that, in the case of no randomization ($p = 0$), in scenario A (typical model) we achieve a higher statistical significance value than expected: with α values around 0.067, we observe a higher chance of type I error than we would expect. For scenario B, conversely, we see α values around 0.05, as we would expect. As we add randomization, the α value tends to get lower, in both scenarios. This is to be expected: by adding randomization in a uniform fashion, we are expecting the two scores distribution to more closely resemble each other, thus decreasing the chance that we would observe difference among the two when there are actually none. In fact, the high values of type I error we observe with no randomization point to the fact that the t-test we are performing is likely to be too powerful, detecting statistical differences when there are none in 5% of the cases. Similar results can observed on the distribution of the α values as we increase N or p, between applying the BT.500 algorithm or opting out of subject screening altogether. However, we can observe that P.910 maintains more stable results even as p increases, in virtue of the randomized subjects being detected and thus removed.

Fig. 9 and 10 depict the variations in achieved power $1 - \beta$ for different outlier detection algorithms, again as a function of number of randomized subjects, and probability of randomization. We can see that, when no randomization is applied ($p = 0$) our simulations for the typical subject model (scenario A) closely approximates the expected power we computed: in Fig. 9, we achieve a power of $1 - \beta = 0.778$ (expected: 0.80), while in Fig. 10, we achieve a power of $1 - \beta = 0.978$ (expected: 0.99). As we progressively increase the randomization in our scores, we see the power steadily decreasing: for an expected power of 0.80, when no outlier detection algorithm is applied, the randomization has a strong effect on the achieved power, getting as low as 0.196 (meaning that in the large majority of the cases, we would not be able to see statistical differences among the two codecs). The



(a) Scenario A: Typical subject model

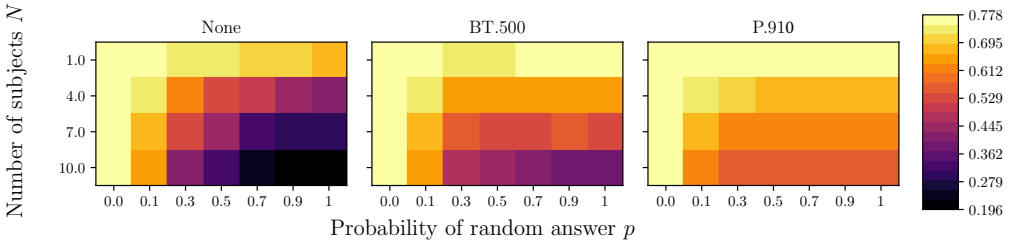


(b) Scenario B: Super-precise subject model

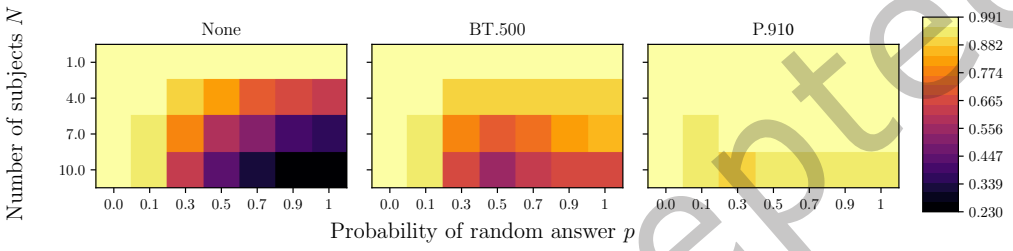
Fig. 8. Observed values of significance level α for different outlier detection scenarios (None, BT.500, P.910), in relation to the probability of random answer (x-axis), and to the number of subjects affected by randomization (y-axis). The difference among the two simulated curves is set to 0.

situation improves when we use the BT.500 outlier detection algorithm: for example, for $N = 4$, we are able to achieve consistent levels of power 0.6, whereas for no outlier detection applied, we would get as low as 0.4. In the worst scenario we consider here, for $N = 10$ and $p = 1$, we have an achieved power of 0.36, almost double to what we achieve when no outlier correction is applied. However, P.910 clearly shows a superior performance: even in the worst-case scenario, we are still achieving a power of 0.6. Similarly, for an expected power of 0.99, P.910 achieves a minimum power of 0.88, whereas BT.500 gets as low as 0.61, and no outlier detection brings the power down to 0.33.

For scenario B (super-precise subject model), we can observe that the achieved power is larger than the expected power, when no randomization is applied: in particular, we achieve a power of 0.99 for an expected power of 0.80, and a power of 1 for an expected power of 0.99. This is in virtue of the parameters of our simulation, which make the two score distributions very distinctive, as little variance is added to the data. Thus, smaller differences among the curves can be detected with the same power. The effect of randomization, however, is equally disadvantageous for scenario B: without performing any outlier detection algorithm, the achieved power lowers down to 0.23 for an expected power of 0.80, and to 0.44 for an expected power of 0.99. Applying outlier detection algorithms clearly improves the achieved power of the test. The best performance is again seen for P.910, whose achieved power for scenario B never goes below 0.9.



(a) Scenario A: Typical subject model



(b) Scenario B: Super-precise subject model

Fig. 9. Observed values of power $1 - \beta$ for different outlier detection scenarios (None, BT.500, P.910), in relation to the probability of random answer (x-axis), and to the number of subjects affected by randomization (y-axis). The effect size is set to 0.064, corresponding to a difference among curves of 0.074. Expected power = 0.80.

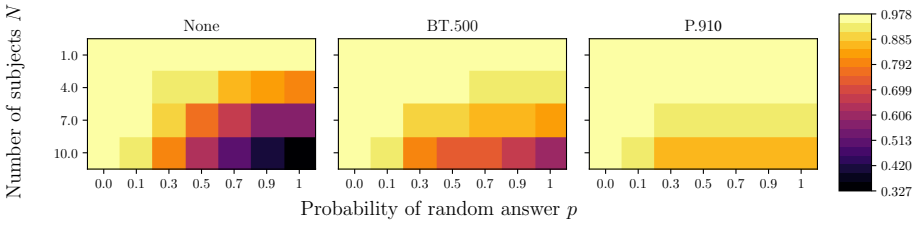
6 Conclusions

In this work, we presented a novel approach to simulate subjective experiments by introducing an algorithm applicable to two distinct scenarios. These simulations enable comprehensive comparisons of different screening methods. More importantly, they provide a foundation for future studies aimed at developing statistical methodologies for subjective experiments and extending the validation or development of new screening methods.

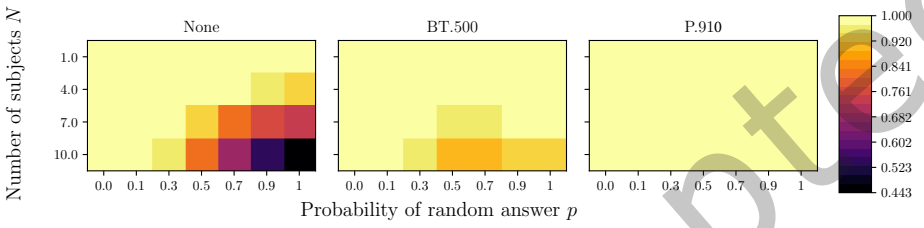
The metrics we proposed for validating subjective experiments offer a deeper understanding of experimental results. They facilitate the evaluation of various setups and algorithms from multiple perspectives, including accuracy, dispersion, and resolution. We anticipate that these metrics will become a standard tool in future simulation studies.

Additionally, the entire codebase for generating all simulations and plots is publicly available at https://github.com/LucjanJanowski/subjective-exp-simulation-subject-screening. We hope that future researchers will leverage this resource in conducting diverse simulations to explore specific aspects of subjective experiments.

The most important conclusion drawn from our simulations is that the kurtosis-based subject screening method described in BT.500 is ineffective under our experimental conditions. It fails to robustly detect random users and consistently underperforms compared to the correlation-based method described in P.910 across various metrics. This conclusion has direct implications for the research community. We strongly advise adopting the P.910 methods—either the correlation-based approach or the AP method, as both are robust in the presence of



(a) Scenario A: Typical subject model



(b) Scenario B: Super-precise subject model

Fig. 10. Observed values of power $1 - \beta$ for different outlier detection scenarios (None, BT.500, P.910), in relation to the probability of random answer (x-axis), and to the number of subjects affected by randomization (y-axis). The effect size is set to 0.098, corresponding to a difference among curves of 0.113. Expected power = 0.99.

strong perturbations. In particular, P.910 might be more suitable for cases in which researchers wish to perform statistical tests that require the entire distribution of scores, such as t-tests or non-parametric tests. Conversely, AP is particularly suitable when researchers do not wish to exclude large swaths of data due to localized errors or outlying behavior, as it allows them to retain the data and weigh it appropriately.

The work here is limited by the outlier and score modeling, which is kept simple to demonstrate our analysis in a baseline scenario. We based our subject model on previous work [13]; however, different models could be consider that account for content or stimulus ambiguity [10]. Based on that, we analysed a simple score scenario with two sources of compression distortion, considering only two subject scenarios (typical and super precise), and a simple outlier scenario with permutation added to a number of subjects with a certain probability. Future work will focus on extending the analysis beyond the scenarios presented in this work to better model different datasets from the literature. Our simulation framework easily allows for such extensions by changing the subject parameters and including different score modeling functions. Our future efforts will include creating more diverse subjective score scenarios (e.g., including more than 2 coding comparisons), different subject models beyond the typical and super-precise case, and diverse outlier models beyond permutation.

Acknowledgment

This work was supported by the Polish Ministry of Science and Higher Education with the subvention funds of the Faculty of Computer Science, Electronics and Telecommunications of AGH University. This work was

partially supported by the European Commission Horizon Europe program, under the grant agreement 101070109, *TRANSMIXR* https://transmixr.eu/. Funded by the European Union.

References

[1] Alexander Berger and Markus Kiefer. 2021. Comparison of different response time outlier exclusion methods: A simulation study. *Frontiers in psychology* 12 (2021), 675558.

[2] Bogdan Ćmiel, Jakub Nawała, Lucjan Janowski, and Krzysztof Rusek. 2023. Generalised score distribution: underdispersed continuation of the beta-binomial distribution. *Statistical Papers* (2023), 1–33.

[3] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. 2009. Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods* 41, 4 (2009), 1149–1160.

[4] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods* 39, 2 (2007), 175–191.

[5] Tobias Hoßfeld, Poul E Heegaard, Martin Varela, Lea Skorin-Kapov, and Markus Fiedler. 2020. From QoS distributions to QoE distributions: A system's perspective. In *2020 6th IEEE Conference on Network Softwarization (NetSoft)*. IEEE, 51–56.

[6] Tobias Hoßfeld, Raimund Schatz, and Sebastian Egger. 2011. SOS: The MOS is not enough!. In *2011 Third International Workshop on Quality of Multimedia Experience*. 131–136. doi:10.1109/QoMEX.2011.6065690

[7] ITU-R. 2023. BT.500 : Methodologies for the subjective assessment of the quality of television images. Geneva, Switzerland. *International Telecommunication Union* (2023).

[8] ITU-T. 2023. P.910: Subjective video quality assessment methods for multimedia applications. Geneva, Switzerland. *International Telecommunication Union* (2023).

[9] Lucjan Janowski, Jakub Nawała, Tobias Hoßfeld, and Michael Seufert. 2023. Experiment Precision Measures and Methods for Experiment Comparisons. In *2023 15th International Conference on Quality of Multimedia Experience (QoMEX)*. 49–54. doi:10.1109/QoMEX58391.2023.10178599

[10] Lucjan Janowski and Margaret Pinson. 2015. The accuracy of subjects in a quality experiment: A theoretical subject model. *IEEE Transactions on Multimedia* 17, 12 (2015), 2210–2224.

[11] Jing Li, Suiyi Ling, Junle Wang, Zhi Li, and Patrick Le Callet. 2020. GPM: A Generic Probabilistic Model to Recover Annotator's Behavior and Ground Truth Labeling. arXiv:2003.00475 [cs.AI] https://arxiv.org/abs/2003.00475

[12] Zhi Li, Christos Bampis, Lucjan Janowski, and Ioannis Katsavounidis. 2020. A Simple Model for Subject Behavior in Subjective Experiments. *Electronic Imaging* 2020 (01 2020), 131–1. doi:10.2352/ISSN.2470-1173.2020.11.HVEI-131

[13] Zhi Li, Christos Bampis, Lucjan Janowski, and Ioannis Katsavounidis. 2020. A Simple Model for Subject Behavior in Subjective Experiments. *Electronic Imaging* 2020 (01 2020), 131–1. doi:10.2352/ISSN.2470-1173.2020.11.HVEI-131

[14] Jakub Nawala, Lucjan Janowski, Bogdan Cmiel, and Krzysztof Rusek. 2020. Describing Subjective Experiment Consistency by P-Value P–P Plot. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 852–861. doi:10.1145/3394171.3413749

[15] Jakub Nawała, Lucjan Janowski, Bogdan Ćmiel, Krzysztof Rusek, and Pablo Pérez. 2022. Generalized Score Distribution: A Two-Parameter Discrete Distribution Accurately Describing Responses From Quality of Experience Subjective Experiments. *IEEE Transactions on Multimedia* (2022), 1–15. doi:10.1109/TMM.2022.3205444

[16] Margaret Pinson et al. 2010. Report on the validation of video quality models for high definition video content. *Video Quality Experts Group* (2010). https://www.its.bldrdoc.gov/vqeg/projects/hdtv/hdtv.aspx

[17] Margaret H. Pinson. 2023. The Precision and Repeatability of Media Quality Comparisons: Measurements and New Statistical Methods. *IEEE Transactions on Broadcasting* 69, 2 (2023), 378–395. doi:10.1109/TBC.2023.3236528

[18] Peter Reichl, Bruno Tuffin, and Raimund Schatz. 2013. Logarithmic laws in service quality perception: where microeconomics meets psychophysics and quality of experience. *Telecommunication Systems* 52 (2013), 587–600.

[19] Emma M Siritzky, Patrick H Cox, Sydni M Nadler, Justin N Grady, Dwight J Kravitz, and Stephen R Mitroff. 2023. Standard experimental paradigm designs and data exclusion practices in cognitive psychology can inadvertently introduce systematic "shadow" biases in participant samples. *Cognitive Research: Principles and Implications* 8, 1 (2023), 66.

Received 31 January 2025; revised 25 July 2025; accepted 7 December 2025