The Efficient Exam

Accepted 22nd August, 2016

ABSTRACT

This paper defines the efficient examination as the one that for a given number of questions maximizes the revealed variability in the distribution of abilities of examinees. Variability is measured by the Gini coefficient which is a robust measure of inequality. Using the properties of the Gini coefficient enables the classification of abilities into different distinguishable types, and differentiating between the concepts of discriminatory power and difficulty of a question. Discriminatory power of a question is defined as the ability of the question to stratify the distribution of examinees according to whether they were able or unable to answer the question, while difficulty is characterized according to whether the ability to answer a specific question increases with the ability demonstrated in the examination. Illustrations of the methodology are presented.

Key words: Latent ability, exam, Gini’s mean difference, Gini’s scores elasticity.

INTRODUCTION

The aim of this paper is to investigate the basic properties of examinations in order to see what can be safely deduced from scores. Ability, even in its simplest form, is a latent variable and does not have a unit of measurement. This property distinguishes ability from other quantitative variables like height or weight of a person. Hence, it is not reasonable to measure it in the same way as quantitative variables, nor that it is appropriate to perform the same analysis as is done with quantitative variables. On the other hand, most of the practical literature dealing with the measurement of ability in a specific field treats scores achieved in examinations as if they are physical quantities like height that can be measured according to a yardstick. The advantage of the latter approach is that one can borrow methods developed to handle quantitative variables in order to apply them on scores achieved in an examination. The downside of this approach is that some of the conclusions derived from the existing literature on measurement of ability can be refuted by an alternative legitimate examination. Not surprisingly, this deviation between the theoreticians advocating the use of variants of Abstract Measurement Theory’, (hereafter AMT), as reflected in Cliff (1977, 1979) and the empirical approach in Lord’s (1980) 'Item Characteristics Theory', (hereafter, ICT) resulted in a fierce debate between advocates of the theoretical purity and the empiricists. Cliff (1992) and Narens and Luce (1993) seem to admit the advantages of the empiricists' arguments, but the debate continues to surface. Recent publications dealing with this debate includes, among others, Bond and Lang (2013a, b), Heene (2013) and Markus and Borsboom (2012).

This paper advocates the substitution of the variance of the scores in the examination by the Gini’s Mean Difference of the scores, (hereafter, GMD or simply Gini) as a method that bridges the gap between the measurement of ability as an ordinal variable, but on the other hand, it includes the decomposition properties of the variance as a special case. It would be fair to admit that the suggestions in this paper have already been suggested in the literature: The AMT advocates suggested using a version of the Gini covariance but they did not translate the theory into a practical tool that can be used as easily as a regression. To support this argument, note that Cliff (1979, p-389, equation (18)), has...
actually developed the equivalent of the covariance in terms of the Gini covariance but did not see that what he has developed is a Gini covariance. Interest in robust regression led Wainer and Thissen (1976) to suggest using the Gini as a substitute for the variance and they developed indirectly a Gini correlation. However, they lacked the computational ability that was later developed. Yitzhaki and Schechtman (2013) present the properties of the Gini as a substitute for the variance and developed the decomposition properties of the GMD. By combining the AMT and ICT approaches, one can avoid presenting conclusions that the AMT approach would reject.

In a nutshell, the properties of the Gini coefficient that enables one to satisfy both approaches are based on the following basic property: Let us consider two random variables: X and Y. Assume that Y is an ordinal variable while X is a quantitative one. For example, let Y be the grades in the examination, while X describes the response to a particular question. Grades are an ordinal variable because by changing the difficulty distribution of the questions in the questionnaire, one may change the grades even if the ranking of the examinees remains the same. Hence, it is not meaningful to present the cumulative distribution of grades since the variate is not well defined. However, the covariance between X and F(Y), where F(Y) is the cumulative distribution of Y is not affected by changing the difficulty distribution of the questions in the questionnaire. This is so because the distribution of all cumulative distributions is uniform, but \( \text{cov} (X, F(Y)) \) is the Gini equivalent of the covariance. The use of the Gini equivalent offers additional information on the underlying distribution as argued by Lambert and Decoster (2005) but on the other hand, limits the use of some practical methods.

The following examples present such a case. There are several international programs like PISA and TIMSS that are intended to compare average achievements of students from different countries and publish them. Those publications are reported on the front page of newspapers and are subject to national debates. Taking into account the latent property of ability, one can point out that in some cases there is an alternative legitimate examination that will change the ranking of countries. Hence, the ranking of groups of students according to average grade is not as robust and conclusive as it sounds.

Since the arguments of the theoreticians concerning the ordinal nature of grades in examinations are ignored by the practitioners, we will replicate them. Unlike quantitative variables that can be measured directly, the way to find out the ability of a person in a specific field is to perform an examination. An examination is a set of questions with different levels of difficulty intended to find out how many questions the person is capable of answering. We argued that since ability does not have a unit of measurement, the method of examination is only useful for ranking examinees according to ability. Would one restrict examinations to only rank individuals, then no problem arises. However, this is only one use of the results of examinations. In order to analyze the results of an examination different kinds of aggregation must be performed. For example, in comparing achievements among countries averages are calculated. Also, in order to construct a proper examination regression techniques are used, for example, the IRT model. Regression methods are based on aggregation. Whenever aggregation is performed it implies that the basic fact that there is no agreeable unit of measurement of ability is ignored.

The fact that in order to analyze the results of the examination one has to aggregate grades implies that one has to impose, explicitly or implicitly units of measurement on a latent ordinal variable. For example, the number of questions each examinee answered correctly can serve as a quantitative variable representing ability. But this is not a true quantitative variable because by changing the difficulty distribution of the questions in the questionnaire the examiner may change its value. We argued that this difference can lead to unreliable conclusions because a change in the assumption concerning the units of measurement may, in some cases, reverse the ranking of groups ranked by the average success of the examinees.

The aim of this paper is to point out the difference between measurement of the latent ability and the measurement of quantitative variables in order to develop the criterion that should be used in measurement of ability. The suggested criterion takes into account the difference between the variables, such that the conclusions derived from measurements of ability can be trusted.

Using the criterion, alternative measures of difficulty of a question based on whether the probability of answering a question increases with the ability demonstrated in the examination and of the discrimination properties of a question in a questionnaire based on the concept of stratification was developed. Discrimination properties of a question are defined as the ability of a question to discriminate among examinees with different ability.

The target of this paper is restricted to the basic and simplest issue: the measurement of uni-dimensional ability (UDA): Person A will be defined as having higher ability than person B in a specific field, if for every question asked, the probability that A will answer correctly is not lower than the probability of B and for at least one question the probability is higher than B. This paper deals with examinations intended to evaluate the distribution of UDA.

### Characteristics of measurements in the field of education

This paper deals with uni-dimensional ability (UDA) only. Since UDA is a latent variable one needs an examination to evaluate it. The purpose of the examination is to unravel the distribution of the ability. To see the difference between unraveling the distribution of a quantitative variable and a
latent one we compared measurement of the distribution of height of a group, with that of UDA. Height can be measured directly and can include random or non-random measurement errors. In contrast, measuring UDA is like measuring the height of examinees that are standing behind a screen. The administrator asks who is taller than 160 cm and each examinee truthfully answers by saying yes or no. The administrator then picks another height and the process is repeated. The "height" of a person would then be determined by the number of questions that were answered positively. An individual's "height" and the average "height" of the examinees would then depend on the "difficulty" of the questions about height. In the case of height, the "difficulty" can be measured by the distance in centimeters between consecutive questions. But since there is no yardstick for measuring the difficulty of a question, difficulty and ability must be assessed by the same scale, the number of questions answered successfully.

Two essential differences between assessments of height and UDA are apparent here: (a) while height is well defined; the criterion for a specific ability is not a-priori clear. That is, the successful answer to a question may be a result for different traits of the examinees. For example, one examinee is able to answer a question in math because of her analytical ability while the other is able to answer the same question because of her superior memory. We refer to this issue as the classification of abilities; (b) while height is a quantitative variable, ability, even UDA, does not have a unit of measurement. As a result, ability must be considered as an ordinal variable. These distinctions play an important role in the proposed methodology.

The implications of (a) is that we need a way of classifying abilities into the different traits that enable answering a question. Classifying abilities can be done by an IRT model. However, the typical IRT model assumes that the distribution of abilities follows a given distribution. YIP (2012) suggested a way of classifying abilities. The approach resembles that delineated by Linacre (1995, 1998). However, this issue is not covered in this paper which is restricted to UDA.

The second difference between ability and height earlier described (b), that is, the issue of quantitative vs. ordinal measurement, resembles the difference between the direct measurement of height (that is, quantitative) and the measurement of the "behind a screen" height (hereafter BS), the latter corresponding to assessment of ability. Consider an examination composed of twenty questions, in which examinees are asked whether their height is higher than 150, 155, 160 cm. BS measurements are equivalent to a particular way of binning (Wainer et al., 2006). That is, all individuals whose height is between 150 and 155 cm, whether they are 151 or 154 cm tall will be classified with the lower height, that is, 150 cm. As such, the continuous variable of ability is converted into a discrete distribution, with increments defined by the difficulty distribution of the questions. However, unlike the example of BS height, there is no quantitative scale for measuring ability. Thus, the assumption of a specific distribution of ability implies converting an ordinal variable of ability into a quantitative one. This implies that non-parametric estimates of ability do not exist, because a distributional assumption, implicit or explicit must be assumed to quantify ability. To attest that, note that the distribution of the grades in an examination is a function of the difficulty distribution of the questions in the questionnaire.

Another way of understanding the implication of the difference between the direct and the BS measurement is to examine the implications of asking about 151 rather than 155 cm. It is as if we would change the step function converting ability into scores of the examination. The more questions that are added in a specific height range, the greater the slope of the transformation converting ability into grades in that range of ability. Adding a question to the examination, or changing the difficulty distribution can be interpreted as applying a monotonic non-decreasing transformation to the function converting the latent ability into grades.

To summarize this, ability is assumed to be a latent continuous variable. An examination is a set of questions intended to assess ability. Since the number of questions in an examination is limited, the examination can be viewed as a transformation from a continuous ability to discrete scores of responses. The transformation depends on the difficulty of the distribution of the questions in the questionnaire. As a result, we are not able to identify the cumulative distribution of ability.

This conclusion can be summarized in the following impossibility theorem:

**It is impossible to estimate the cumulative distribution of ability:** To prove it, all we need is to serve two examinations to the same group of examinees. One examination will be composed of one easy question and the rest will be difficult questions, while the other examination includes many easy questions and one difficult question. One distribution of grades will be positively skewed while the other will be negatively skewed.

**The purpose of the examination**

An examination is a series of N questions that is intended to describe the distribution of ability in a specific field of knowledge. One can classify examination into two groups according to the purpose of the examination: Acceptance and grading examinations.

**An acceptance examination:** This is an exam intended to classify the examinees into two groups, those that are accepted into a program and those that are rejected.

**A grading examination:** This is an examination with the purpose to reveal the entire distribution of ability among all
examinees. The target functions of the two types of examinations may be different. In this paper, only grading examination was dealt with.

It is clear that the number of questions determines the ability of the examination to describe the distribution of ability. For a population with \( H \) examinees, one needs \( H \) questions to identify the ability of each examinee. An examination with \( N < H \) questions can only classify the between-group variability of ability among the examinees, because binning is unavoidable.

In summary, it is assumed that the distribution of ability among examinees is given but unknown. The examination can only estimate the between-group variability. Therefore, the best examination is the one that can find out most of the variability, that is, the maximum between-group variability. This leads us to characterization of the efficient examination.

Characterization of the efficient exam

The most efficient examination is defined as the one that for a given number of questions, \( N \), minimizes the difference between the variability of the distribution as revealed by the examination and the true variability among the examinees. However, the data we have from an examination is the ranking of examinees such that we can use the property that all cumulative distributions are uniformly distributed between \([0, 1]\). The characterization of the efficient exam was based on this property.

For this purpose, we assumed the following initial assumptions:

(a) The examination is prepared before it is conducted: This assumption rules out examinations that are based on knowing the response to each question before asking additional questions, as is the case in an oral examination.

(b) There is no randomness in the response (to be modified later): That is, for each level of ability, \( a \), there is a level of difficulty \( d(a) \) such that:

\[
p(a, d) = \begin{cases} 
0 & \text{for } d > d(a) \\
1 & \text{for } d \leq d(a) 
\end{cases} \tag{1}
\]

Assumption (1) implies that the probability of answering a question is either zero or one, depending on the ability of the examinee and the difficulty of the question. As a result we have at most \( N+1 \) observations on the cumulative distribution of ability of the examinees.

At this stage, the difficulty of a question is defined according to the fraction of correct responses, that is, the number of correct answers divided by the number of examinees. Arranging the questions in an increasing order of difficulty we get that the least able examinee was able to correctly answer from zero to \( n_1 \) questions, the second worst was able to answer all the questions that the least able answered and additional questions ranging from zero to \( n_2 \) questions, etc. As a result, we get up to \( N \) observations on the true cumulative distribution of ability. That is:

\[
\sum_{i=1}^{N} n_i = N \tag{2}
\]

The issue we are dealing with is what should be the distribution of \( n_1, n_2, ..., n_N, n_{N+1} \) of the most efficient examination. The most efficient examination maximizes the between-group variability of the \( N \) observations. Maximizing between-group variability minimizes the intra-group variability.

Let \( f \) be the proportion of the examinees that were able to answer \( \sum_{i=1}^{N} n_i \) questions of the exam and failed to answer \( H \) the rest and Let \( c = \frac{100\% (N + 1)}{N} \), be the equal percentage of examinees per interval.

Proposition 1

(a) The efficient examination is characterized by \( f_i = i \times c \) for \( i=0, 1, ..., N \). That is, the efficient examination will utilize all the questions of the questionnaire and divides the cumulative distribution of the examinees into portions of equal size of examinees.

(b) Provided that the questions have equal weights, then, the distribution of the grades in the efficient examination converges to the uniform distribution as the number of questions increases.

Before we proved the proposition it is worth to describe the properties of the Lorenz curve that will be used in order to derive a geometric proof.

Properties of the Lorenz Curve

The Lorenz curve is known since Lorenz (1905). Its properties are covered in numerous books, among them Yitzhaki and Schechtman (2013: 75-98) also covers the Absolute Lorenz curve and the relationship between the curves, the Gini coefficient and the coefficient of variation. Figure 1 presents a typical Lorenz curve. On the horizontal axis, the cumulative distribution of the examinees according to ability is presented, while the vertical axis presents the share of the least able examinees in the total scores in the examination. The properties used in the proof of the proposition are the:

- The Lorenz curve starts at \((0,0)\) and ends up at \((1,1)\);
- The Lorenz curve is monotonically increasing and convex toward the horizontal axis;
Academia Journal of Educational Research; Yitzhaki. 145

Figure 1: A) typical Lorenz curve.

- The line of Equality (LOE) is the 45° line. It represents the Lorenz curve of examinees with identical ability;
- The Gini coefficient is twice the area enclosed between the LOE and the Lorenz curve;
- The between-group component of the Gini coefficient is twice the area enclosed by linear segments connecting the points A, B, C, that are on the Lorenz curve and the diagonal. Figure 1 presents a hypothetical Lorenz curve and the between-group Lorenz curve.

Proof of proposition 1

The proof of the first part is composed of two stages:

a) That \( f_i > f_{i-1} \) for \( i = 1, \ldots, N \);

b) That \( f_i - f_{i-1} = \) constant for \( i = 1, \ldots, N \).

The implication of (a) is that no question should be wasted; the implication of (b) is that the efficient exam spreads all the questions evenly along the cumulative distribution of ability.

Proof of a: This is trivial. It is composed of substituting a straight line that adds zero to the area of between-group Gini by a triangle with a positive area that increases between-group Gini.

Proof of b: Assume that two intervals on the horizontal axis are not equally spaced. They produce a triangle in the between-group Gini. Due to the convexity of the Lorenz curve, one can increase the area of the triangle by moving to the point where the slope of the Lorenz curve is equal to the slope of the base. Since the curve can have any convex shape, every point that is not in the middle can have an alternative mirror image that makes it an inefficient point. Hence, the only point without another point that cancels its optimality is the middle point.

The proof that the distribution of grades in the efficient examination converges to the uniform as the number of questions increases is trivial and it becomes clear in the next section. QED.

Proposition 1 implies that because we do not know the distribution of ability, we can use the fact that all cumulative distributions are uniformly distributed between zero and one. The most efficient examination is the one that enables us to uniformly divide the range that we are interested in grading it. This will maximize the between-group component that we can reveal over the range of
ability. Would we know the distribution of ability we could increase the between-group Gini, but we do not have this information. We therefore examined the assumptions that led to it. Assumption (a), that the examination is prepared before it is conducted is a must for all written examinations. Assumption (b) that there is no randomness in the response is a strong one. We may substitute it by adding a random element in the response that is statistically independent and has an equal distribution for all questions asked and levels of ability. This will complicate the proof, forcing us to move to averages and expected values, but the logic will be similar.

EMPIRICAL ILLUSTRATIONS

Two illustrations are presented here. Both are based on the psychometric entrance test (PET) conducted by the National Institute for Testing and Evaluation in Israel. The scores are those of all participants in the examinations conducted in the academic year 1999 to 2000 and fulfilled the basic requirements for possible acceptance to a university. Although, the examination is called an acceptance (that is, entrance) examination, it is an entrance examination to almost all institutes of higher education and different departments set different entrance requirements. Hence, the test fits the required properties of grading exam. The first illustration is based on the examination in Mathematics. The range of raw scores is (60 and 150). The number of observations is 17,867. Figure 2 presents the density of the distribution of the grades. The horizontal axis presents the grades while the vertical axis presents the percentage of the examinees in each grade. Note that the density is not smooth. This can be a result of either random fluctuations, or that there is “bunching” of questions in the questionnaire such that the ability to answer a specific question is correlated with the ability to answer another question. However, those issues are beyond the scope of this paper.

The efficient examination would have gotten a uniform distribution. As can be seen this is not the case of the density function in Figure 2 and more questions in the middle of the distribution are required. Since our purpose is to illustrate a point there is no point in conducting a formal statistical test because it is clear that subject to the size of the population, the uniform distribution hypothesis would have been rejected.

The second illustration is based on the test in English conducted in the same year. Figure 3 presents the density function of this examination. As can be seen the density is flatter than the density of grades in examination of mathematics such that we can conclude that the examination in English is more efficient than the examination in mathematic. On top of that, the deviation from a density of a uniform distribution is at the extreme
points, such that one can claim that the lower range is not relevant since the examinees will not be accepted in any department, while the extreme right hand is not relevant because the examinees will be accepted by all departments. Therefore, we may conclude that the examination in English is more efficient than the examination in mathematics. However, it is worth to add a disclaimer that it may be easier to write an efficient examination in English than in mathematics since the subject matters differ and the random errors may be different.

THE DISCRIMINATORY POWER OF A QUESTION

This section intends to suggest a measure of the discriminatory power of questions. The ideal question will have the following property: All those identified with high ability will be able to come up with the correct answer and all those with low ability will not be able to answer it. This resembles the sociological concept of stratification. Lasswell's (1965: 10) defines stratification as: "in its general meaning, a stratum is a horizontal layer, usually thought of as between, above or below other such layers or strata. Stratification is the process of forming observable layers, or the state of being comprised of layers". The ideal question stratifies the examinees into two groups: those who are able to answer and those who are not.

Properties of the Gini coefficient were used in order to evaluate the discriminatory power of a question in a questionnaire. The discriminatory power is the probability that a randomly selected member of the group who failed to answer the question also got a low grade in the examination. As far as we know, this issue has not been investigated before. The methodology that is useful is a special case of ANOGI (Frick et al., 2006), which is similar to ANOVA except that it is applied to the Gini coefficient instead of the variance. Actually, we are not interested in the Gini coefficient because as we have argued before, we do not want to rely on the properties of the grades of the examination; rather, we are interested in one component of the decomposition called stratification.

In analyzing the discriminatory power of a given question, we divided the examinees into two groups: Those who successfully answered the question (denoted by subscript a) and those who failed to answer (denoted by subscript p representing poor grades). In an imaginary world with no errors or other factors, we expect the groups to form perfect strata of the examinees in the distribution of grades of the examination. To see this, note that assumption (b) implies that in a perfect case with appropriate questions and no random errors, for each question, all examinees up to a given ability are unable to answer a question, while
those above this threshold ability can answer the question. This gives us the connection between stratification and discrimination properties of a question. The structure of the section is the following: first we introduced ANOGI, the analysis Of Gini, which is similar to ANOVA, except that it also includes a parameter representing stratification. The presentation of ANOGI is based on the general case in which there are more than two groups. The general case can be applied to a question with several levels of responses. Then, we replicated ANOGI for the case of the two groups, which is typical to many examinations, such that we can get additional information on the properties of the stratification parameter.

ANOGI - the general case

Let \( Y_i, F_i(Y), f_i(Y), \mu_i, G_i \) and \( p_i \) represent the grade (variante), the cumulative distribution, density function, expected value, Gini coefficient and the share of sub-population \( i \) in the overall population of examinees, respectively \( (i=1,...,n) \). The Gini coefficient of group \( i \) is:

\[
G_i = \frac{2 \text{cov}_i(Y, F_i(Y))}{\mu_i},
\]

which is twice the covariance between the score \( Y \) and the rank \( F_i(Y) \), standardized by the mean score \( \mu_i \) and \( \text{cov}_i() \) is the covariance calculated under the distribution \( F_i \).

Let \( s_i = p_i \mu_i / \mu_u \) denote the share of group \( i \) in the overall scores. Then, the following formula holds:

\[
G_u = \sum_{i=1}^{n} s_i G_i + \sum_{i=1}^{n} s_i (G_i(1) - 1) + G_BP + (G_B - G_BP)
\]

The components of (4) are hereby highlighted.

The overlapping parameter

Overlapping can be interpreted as the inverse of stratification. Stratification is a concept used by sociologists. One can rarely find a perfect stratification. Therefore, an index which quantifies the degree of stratification is called for. The index of overlapping (to be defined below) quantifies the extent to which the different sub-populations are stratified in the overall distribution. Formally, overlapping of the overall population by sub-population \( i \) is defined as:

\[
O_i = O_{ui} = \frac{\text{cov}_i(Y, F_u(Y))}{\text{cov}_i(Y, F_i(Y))},
\]

Where \( \text{cov}_i \) is the covariance according to distribution \( i \). (For convenience, the index \( u \) is sometimes omitted).

The overlapping index (5) can be further decomposed to identify the overlapping of sub-population \( i \) with all other sub-populations that comprise the union. This further decomposition of \( O_i \) is:

\[
O_i = \sum_{j \neq i} p_j O_{ji} = p_i O_{ui} + \sum_{j \neq i} p_j O_{ji} = p_i + \sum_{j \neq i} p_j O_{ji}
\]

Where \( p_i \) is the share of sub-population \( j \) in the union and \( O_{ji} = \frac{\text{cov}_i(Y, F_j(Y))}{\text{cov}_i(Y, F_i(Y))} \) is the overlapping of group \( j \) by group \( i \). The properties of the overlapping index \( O_{ji} \) are the following:

(a) \( O_{ji} \geq 0 \). The index is equal to zero if no member of the \( j \)-th distribution lies in the range of distribution \( i \). (that is, group \( i \) is a perfect stratum);
(b) \( O_{ji} \) is an increasing function of the fraction of population \( j \) that is located in the range of population \( i \);
(c) For a given fraction of distribution \( j \) that is in the range of distribution \( i \), the closer the observations belonging to \( j \) are to the expected value of distribution \( i \), the higher the \( O_{ji} \);
(d) If the distribution of group \( j \) is identical to the distribution of group \( i \), then \( O_{ji} = 1 \). Note that by definition \( O_{ii} = 1 \). This result explains the second equality in (6). Using Equation (6) it is easy to see that \( O_{ji} \geq p_{ij} \), which is a result to be borne in mind when comparing different overlapping indices of groups with different sizes;
(e) \( O_{ji} \geq 2 \). That is, \( O_{ji} \) is bounded from above by 2. This maximum value is reached if all observations belonging to distribution \( j \) that are located in the range of distribution \( i \) are concentrated at the mean of distribution \( i \). Note, however, that if distribution \( i \) is known then it may be that the upper bound is lower than 2 (Schechtman, 2005);
(f) In general, the higher the overlapping index \( O_{ji} \), the lower the \( O_{ij} \). That is, the more group \( j \) is included in the range of group \( i \), the less group \( i \) is expected to be included in the range of group \( j \).

Properties (a) to (f) showed that \( O_{ji} \) is an index that measures the extent to which sub-population \( j \) is included (overlapped) in the range of sub-population \( i \). Note that the indices \( O_{ji} \) and \( O_{ij} \) are not inter-related by a simple relationship. However, it is clear that the two indices of overlapping are not independent.

The components of (4) are hereby highlighted.

Between-groups component \( G_B \) and its properties

There are two parameters that represent the between group Gini in the decomposition of the Gini of the overall
population. One is defined in Yitzhaki and Lerman (1991) denoted by $G_B$ and the other by Pyatt (1976) denoted by $G_{BP}$, the difference between them plays a major role in the concept of stratification. The between-groups inequality $G_B$ is defined in Yitzhaki and Lerman (1991) as:

$$G_B = \frac{2 \text{cov}_u(\mu, \bar{F}_u)}{\mu_u} \quad (7)$$

$G_B$ is twice the covariance between the mean grades of the sub-populations and the sub-populations’ mean ranks in the overall population, divided by overall mean grade. That is, each sub-population is represented by its mean grade and by the mean rank of its members in the overall distribution. The term $G_B$ equals zero if either the mean grades or the mean ranks are equal for all sub-populations. In extreme cases $G_B$ can be negative, which occurs when the mean grades are negatively correlated with the mean ranks of the members of the sub populations. For example, imagine a case where in one sub population there are a few extremely good examinees while the others are extremely bad. Then, the average grade of the group will be relatively large while the average rank of its members will be relatively small.

$G_B$ is not a pure Gini coefficient because $\bar{F}_u$ is not the cumulative distribution of the variable $\mu_u$. An alternative between-group Gini ($G_{BP}$) is defined by Pyatt (1976) given as:

$$G_{BP} = \frac{2 \text{cov}_B(\mu, F_u(\mu))}{\mu_u} \quad (8)$$

In this definition, the between-group Gini is based on the covariance between the mean grade in each sub-population and its rank among the mean grades of the overall population. This between-groups component is a pure Gini (of the vector of means). The difference between the two definitions is in the rank that is used to represent the group: under Pyatt’s approach it is the rank of the mean grade of the sub-population, while under Yitzhaki-Lerman it is the mean rank of all members belonging to the sub-population.

It can be shown that:

$$G_B \geq G_{BP} \quad (9)$$

The upper limit of $G_B$ is reached and Equation (9) holds as an equality when the groups occupy non-overlapping ranges of grades (that is, perfect stratification) because in that case the ranking of the means of the groups is identical to the average of the rankings of the individual members of the groups. Therefore, the difference between the two can supply an indication of the quality of the overlapping.

**The case of two groups**

In many examination the examinee either answers the question correctly or fails to answer. This means that for each question, there are two groups: those who failed, who are expected to also have poor grades in the examination and those who answered are expected to have high grades. Monti and Santoro (2011, hereafter MS) performed ANOGI for the case of two groups, getting new insights that are extremely useful. The results that we rely on in this paper are: (a) they prove that the ratio of the two between-group Ginis is an index of stratification. (b) they interpret the ratio of between-group components as representing the probability that a random member of the failing group (on average) to get higher scores than a random member drawn from the (on average) higher group. We refer to this property as the discriminatory power of a question. It yields a quantitative answer as to what degree the conclusion reached by comparing the mean grades of the groups also applies to members of the groups. Following are the main results of MS (2011).

Let $\mu_a$ and $\mu_p$ denote the average scores in the exam of each group, respectively. To keep up with the notation used in MS denote the size of the groups of examinees by $n_a$ and $n_p$. In the case of UDA exam we expect that $\mu_a > \mu_p$. Otherwise, the method used in classifying abilities would classify the ability needed to answer the question as different from the ability to answer the questionnaire. MS (2011) defined an index which is the ratio of the two between-group Gini:

$$I = \frac{G_B}{G_{BP}} \quad (10)$$

MS (2011) proved that $I$ is equal to (equation 9: 416):

$$I = 1 - \frac{2 \sum_{h=1}^{n_p} (R_{ph} - r_{ph})}{n_a n_p} \quad (11)$$

Where $R_{ph}$ is the rank of examinee $h$ belonging to the group who failed to answer the question, in the overall distribution of examinees, while $r_{ph}$ is the rank of the same examinee in the distribution of those who failed to answer the question.

**The properties of the index $I$**

The maximum value of $I$ occurs when there is perfect stratification. In this case, all members of the failing to
answer group also get the lower grades in the examination such that \( R_{ph} = r_{ph} \) for all \( h \) and the maximum value of \( I = 1 \) is reached. In this case, \( \text{Prob}\ (Y_p > Y_d) = 0 \), where \( \text{Prob}\ () \) is the probability of transvariation, that is, the probability that a randomly chosen member of the failing to answer group will get a higher grade than a randomly selected member of the group who answered the question is zero. The discriminatory power of the question is perfect, that is, assumption 4.1 is observed by the data. However, this is not a typical case and because of random errors we should expect a lower discriminatory power.

In general, the more examinees that belong to the group that successfully answered the question are in the range of the failing to answer group and the higher the difference \( (R_{ph} - r_{ph}) \) the lower is \( I \).

The minimum value of \( I \) that is relevant to is \( I = 0 \). In this case, MS showed that \( \text{Prob}\ (Y_p > Y_d) = 0.5 \). In this case, the ranking of groups according to the mean grades in answering the question has no discriminatory power over the members of the groups.

The index \( I \) can be negative, but then according to YIP (2012) the question does not belong to the subject matter of the examination.

The index \( I \) can be interpreted as providing a quantitative evaluation of the quality of the information contained in the responses to the question: If \( I = 1 \), that is, if we detect full stratification, then, the conclusions reached from the mean response apply to every member of the two groups. If \( I = 0 \), then, the conclusions reached from the comparison of the means does not apply to comparisons of individual examinees since the probability that a randomly chosen of those who failed to answer the question to get a higher grade in the examination from a randomly chosen member of those who answered the question is 0.5, which means no informational content at all.

Finally, it is worth to explain the difference between monotonic correlation, which is the concept suggested in YIP (2012) to distinguish between two traits that affect the success in an examination and the discrimination power of a question as suggested in this paper. Let \( X, Y \) be the binary response to a question and the scores in the examination respectively. A sufficient condition for being able to distinguish whether the response and the examination belong to different abilities, we need that \( E(X|Y=y) \) is not monotonically increasing in \( y \). If \( E(X|Y=y) \) is an increasing function of \( y \) the subject matter of \( X \) is indistinguishable from the subject matter of \( Y \). Stratification, on the other hand, answers a question that follows the determination that a question belongs to the subject matter of the examination. It answers the question whether all those who failed to answer a specific question also fail to get high scores in the examination. For a question to be a good discriminator, the correct answers of \( X \) should be clustered at the top distribution of \( Y \). The distinction can be sometimes obscure because if \( X \) is a perfect discriminator then it is also belonging to the subject of \( Y \). This means that there should be a hierarchical order: first one has to determine whether a given question belongs to the subject matter of the examination and only the questions that are good discriminators be chosen. In some sense this is equivalent to the existing procedure of applying TestGraf (Ramsay, 1991, 2000) before estimating the ICC.

### Table 1: The data in the artificial example.

<table>
<thead>
<tr>
<th>Examinee Column number</th>
<th>Score</th>
<th>Q1</th>
<th>Q2</th>
<th>F(Exam)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>0</td>
<td>1</td>
<td>0.3</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>0</td>
<td>1</td>
<td>0.4</td>
</tr>
<tr>
<td>5</td>
<td>60</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>6</td>
<td>70</td>
<td>0</td>
<td>0</td>
<td>0.6</td>
</tr>
<tr>
<td>7</td>
<td>90</td>
<td>1</td>
<td>1</td>
<td>0.7</td>
</tr>
<tr>
<td>8</td>
<td>91</td>
<td>1</td>
<td>0</td>
<td>0.8</td>
</tr>
<tr>
<td>9</td>
<td>92</td>
<td>1</td>
<td>1</td>
<td>0.9</td>
</tr>
<tr>
<td>10</td>
<td>93</td>
<td>1</td>
<td>1</td>
<td>1.0</td>
</tr>
</tbody>
</table>

An illustration

The purpose of this is to illustrate the calculation of the stratification index. To avoid the issue of whether the question belongs to the subject matter of the examination, an issue that is dealt with in YIP (2012), we presented an illustration that is based on an artificial example.

Table 1 presents the artificial data concerning the grades in the examination and the responses to two questions: The first column on the left is a number assigned to each examinee while the second column represents the grade achieved in the examination in columns 3 and 4 and 1 represents a correct answer while 0 indicates a failure to answer and columns 5 represents the cumulative distributions of the grades in the examination. The difficulty levels of the two questions are equal because each
question is correctly answered by 50% of the examinees. The average grade for examinees is 61.1. Table 2 presents the calculations for comparing stratification of Q1 vs. Q2; The ratio of I=G/B for Q1 is 0.92 while the same ratio for Q2 is 0.40 such that Q1 is a better discriminator than Q2.

### The difficulty of a question

The usual way of ranking questions according to the level of difficulty is by the proportion of examinees that are able to answer the question. This may be misleading if, for example, those who were able to answer the question are the less able examinees. Since the examination offers a ranking of examinees according to the trait examined it seems reasonable to define the difficulty of a question according to whether the proportion of those who answer it increases with the ability demonstrated in the examination. This problem is similar to the problem of defining the degree of progressivity of taxes, where we are interested to see whether the tax is borne by the high ability persons. Therefore, we can borrow the methodology used to define progressivity and apply it with some modifications in the terminology into classifying questions according to difficulty level. The parameter of classifying progressivity of taxes is called Gini’s Income elasticity, GIE and the main modification is by replacing income by scores in the examination. Note, however, that this condition is sufficient but not necessary.

We are interested in the derivative of the Gini coefficient of grades with respect to an increase in the weight of the kth question. Let w_k be the weight attached to the kth question in calculating the score in the examination. Let w_k(ε) = w_k(1+ ε), where ϵ > 0, and ε→0, that is, ε is a small positive term converging to zero. Since the sum of the weights equals one, an increase in the weight of a question is compensated by a proportional decline in the weights of the rest of the questions. Let η_k be the GSE of the k-th question. Then:

\[ \eta_k = \frac{\text{cov}(x_k, F(g)) \mu_g}{\text{cov}(g, F(g))} \mu_k = \beta_{N \times H} \frac{\mu_g}{\mu_k} \]

Where \( \mu_k \) and \( \mu_g \) represent the average success in answering question k and in the examination, respectively, and \( \beta_{N \times H} \) is the Gini non-parametric regression coefficient of responses to question k on the grades in the examination. The properties of the GSE are the following:

1. If \( \eta_k < 0 \), it means that the higher the grades the lower the success in answering the question. It implies that the question does not belong to the subject matter of the examination. Note, however, that this condition is sufficient but not necessary.
2. If \( \eta_k = 0 \), it means that the number of examinees that answer the question does not change along the grades. This may happen if the ability to answer the question is statistically independent of the grades in the examination. This is an additional case indicating that the question does not belong to the subject matter of the examination.
3. If \( 0 < \eta_k < 1 \), implies that the proportion of examinees who answer the question increases with an increase in the grades. It indicates a relatively easy question increase.
4. If \( \eta_k = 1 \), implies that the percentage of those who answer the question is constant along the distribution of grades.
5. If \( \eta_k > 1 \), implies that the higher the grade the higher the percentage who answered the question. However, unlike the case of taxation, the weighted sum of the GSE is not equal to one. The explanation to this divergence is that increasing all the weights by a given proportion increases the relative inequality since in the efficient examination the group not able to answer any question does not benefit from the increase in score.

To sum it up, the greater the GSE the more difficult is the question because those who manage to answer correctly are at the top of the grades of the examination.

### Table 2: Main parameters.

<table>
<thead>
<tr>
<th>Variable</th>
<th>All</th>
<th>Q1</th>
<th>Q2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average scores of failed</td>
<td>37</td>
<td>45.2</td>
<td></td>
</tr>
<tr>
<td>Average scores of Answered</td>
<td>87.2</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>Average Fp</td>
<td>0.3</td>
<td>0.32</td>
<td>0.4</td>
</tr>
<tr>
<td>Average Fa</td>
<td>0.8</td>
<td>0.78</td>
<td>0.7</td>
</tr>
<tr>
<td>G_B</td>
<td>0.18</td>
<td>0.078</td>
<td></td>
</tr>
<tr>
<td>G_BP</td>
<td>0.197</td>
<td>0.197</td>
<td></td>
</tr>
</tbody>
</table>
Table 3: The Size of an Efficient Exam and the Gini coefficient.

<table>
<thead>
<tr>
<th>Number of questions</th>
<th>Gini coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>0.44</td>
</tr>
<tr>
<td>10</td>
<td>0.36</td>
</tr>
<tr>
<td>100</td>
<td>0.34</td>
</tr>
<tr>
<td>10000</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Illustrations of the properties of the efficient exam

Efficient examination

Consider a given latent distribution of abilities and an examination with N questions. The most efficient ideal examination divides the group of examinees into N+1 groups of equal size, with the least able group not able to answer any question, while the highest ability group answers all the questions.

To make the calculation easy to follow, we assumed that the number of examinees is \( H = (N + 1) \times c \) where c is the size of each group. To simplify the calculations, we assumed that c = 1. That is, in each group includes one person. This simplification does not affect any result of this section. The efficient examination is characterized by a uniform distribution of scores. One can use this additional information to get the following additional results:

We illustrated the additional results by presenting an example of efficient examination based on ten questions with binary scores. That is the response to each question is either true or false. The weight attached to each correct answer is equal and the range of scores is between [0, 100]. Hence, each correct answer yields \( w = \frac{100}{N} \). That is, the scores are (0, w, 2w, ..., Nw) . Since w is a constant, it will be ignored (Paul Allanson, 2013).

Following some tedious algebra, the following results are obtained:

(a) The value of the numerator of the Gini coefficient is given as:

\[
\text{Cov}(Y, F(Y)) = \frac{N(N+2)}{12(N+1)}
\]  

(13)

Where \( Y \) represents the grades in the examination. The Gini index of inequality of the scores of the examination is given as:

\[
G(Y) = \frac{2\text{cov}(YF(Y))}{\bar{Y}} = \frac{(N+2)}{3(N+1)}
\]  

(14)

Where \( \bar{Y} \) is mean scores. As can be seen, the Gini coefficient of scores declines as the number of questions increases.

This is an unexpected result and hence deserves a proposition.

**Proposition:** The bigger the number of questions in the exam, the smaller the Gini coefficient of scores among examinees in the efficient exam.

**Proof:** See Equation (9.2).

Table 3 presents the values of the Gini coefficients of scores for several values of the number of questions in the examination. As can be seen, the upper bound of the Gini of the efficient examination is 0.5 (for an examination composed of one question), and the lower bound is 0.333 for an examination with 10000 questions. The main conclusion is that the number of questions determines the inequality among the examinees. The explanation to this result is that the more questions are in the examination the closer the distribution of scores to the uniform distribution.

The difficulty of a question

We now present the concept of the "difficulty" level of a question. The usual way is to determine the difficulty of a question by the number of examinees who were able to answer it. This definition of difficulty ignores the fact that we already have a measure of the specific ability examined in the examination, which is the distribution of the scores in the examination. We should expect that the more difficult the question is the lesser the ability of examinees to answer questions. Alternatively, to rely on the specific ability examined we should expect that the proportion of examinees that are able to answer a specific question increases as the scores in the examination increase.

To illustrate the estimation of difficulties of questions we consider an efficient examination with ten questions. The scores in the examination are composed of eleven groups of scores, according to the number of questions that are answered (Table 4). The first column presents the number of questions answered, the second presents the scores in the examination, while the rest present whether the response was correct (1) or incorrect (0) for each question. Another interesting point that may be viewed as a puzzle is that since the GSE is based on the Gini coefficient one should expect that the weighted sum of the GSE, weighted by the weight attached to each question will be equal to...
Table 4: The raw (artificial) data in the efficient examination.

<table>
<thead>
<tr>
<th>Question answered / No. of questions answered</th>
<th>Score</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
<th>Q5</th>
<th>Q6</th>
<th>Q7</th>
<th>Q8</th>
<th>Q9</th>
<th>Q10</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>30</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>50</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>60</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>70</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>80</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>90</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5: Summary Statistics for the efficient examination.

<table>
<thead>
<tr>
<th>Question number</th>
<th>Average score</th>
<th>Gini Regression coefficient*</th>
<th>Ratio of Averages</th>
<th>GS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>10</td>
<td>0.045</td>
<td>0.18</td>
<td>0.25</td>
</tr>
<tr>
<td>Q2</td>
<td>9</td>
<td>0.081</td>
<td>0.036</td>
<td>0.5</td>
</tr>
<tr>
<td>Q3</td>
<td>8</td>
<td>0.11</td>
<td>0.054</td>
<td>0.75</td>
</tr>
<tr>
<td>Q4</td>
<td>7</td>
<td>0.127</td>
<td>0.072</td>
<td>1.0</td>
</tr>
<tr>
<td>Q5</td>
<td>6</td>
<td>0.136</td>
<td>0.09</td>
<td>1.25</td>
</tr>
<tr>
<td>Q6</td>
<td>5</td>
<td>0.136</td>
<td>0.11</td>
<td>1.5</td>
</tr>
<tr>
<td>Q7</td>
<td>4</td>
<td>0.127</td>
<td>0.127</td>
<td>1.75</td>
</tr>
<tr>
<td>Q8</td>
<td>3</td>
<td>0.11</td>
<td>0.145</td>
<td>2.0</td>
</tr>
<tr>
<td>Q9</td>
<td>2</td>
<td>0.081</td>
<td>0.16</td>
<td>2.25</td>
</tr>
<tr>
<td>Q10</td>
<td>1</td>
<td>0.045</td>
<td>0.14</td>
<td>2.5</td>
</tr>
<tr>
<td>Exam</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Since the distance between scores is equal, the numerical value of the Gini regression coefficient is equal to the Ordinary Least Squares regression coefficient.

Zero. This is because we expect that raising all the weights by a given percentage will have an inflationary effect on the scores but will not change the relative measure of variability. The explanation to this deviation from what is expected is the existence of the group of examinees who are not able to answer any question so that they are not benefitting from the increase in the weight attached to each question, so that raising the weights attached to all questions increases inequality among the examinees.

Table 5 presents the summary statistics of the efficient examination. The first column presents the question, the second - the number of examinees who answered correctly, the third presents the Gini regression coefficient, the forth the ratio of averages of answers to the question relative by the average of the examination, while the fifth presents the GSE. From the table we can see the following: Neither the Gini nor the OLS regression coefficients can serve as indicators of the difficulty of the question since they are symmetric with respect to the difficulty of a question. That is a difficult question gets the same regression coefficient as its symmetric easy question. This is in contradiction to the practical use of the regression coefficient as an indicator of the difficulty of the question. Finally, the discrimination property of each question in the efficient examination is perfect because the ranking of the average score for each question is identical to the average ranks of the answers. Hence, the efficient examination cannot be used to illustrate the discriminatory power of a question.

CONCLUSION AND RECOMMENDATION

The aim of this paper is to define the efficient examination as a criterion for evaluating examinations and to characterize it based on the solid data that one gets from an examination without imposing assumptions on the distribution of abilities that cannot be verified or tested. The efficient examination is defined as the one which maximizes the revealed (between-group) variability of ability, subject to the constraint that the number of questions in the examination is limited. It is characterized ex-post as the one that results in a uniform distribution of grades among the examinees. Having defined the efficient exam we have presented and illustrated the discrimination and the difficulty properties of a question.

The use of the Gini enables the efficient examination to be a compromise that can satisfy both the theoretical approach
and the empirical one, because it uses the concepts of the empiricists without violating the assumptions required for the analytical one.

Further research is needed in order to fully utilize the additional properties of the Gini that can improve our ability to find out questions that discriminate well among the examinees. For example, MS (2011) developed graphical presentation of stratification. Also, user friendly software will help improve the suggested methodology.

ACKNOWLEDGEMENT

I am grateful to Ingram Olkin, Maria Giovanna Monti and Yolanda Golan for assistance and valuable suggestions in the course of this study.

REFERENCES