嘉大教育研究學刊，第 30 期：53-78
2013 國立嘉義大學　教育學系

教育公平性測量：Gini 係數衍生指標的效能分析

李茂能[*]

摘　要

本模擬研究旨在探討 11 種 Gini 係數導向的相對效能與 X 軸等份分割數、教育資料分配型態及離散程度屬性間之關係。一般而言，教育 Gini 係數的相對效能會受到資料的組別數、離散程度與分配型態所支配。教育 Gini 係數的測量正確性會，隨著組別數的增加或離散程度的降低而改善。不管資料的組別數、離散程度與評鑑效標為何，G1 與 G2 指標在資料出現正偏時會產生最大的估計誤差，在資料出現負偏時會產生最小的估計誤差。 其它的教育 Guni 指標 G3, G4, G5, G6, G7 僅在組別數為 5 且資料出現正偏時，會出現類似最大估計誤差；在組別數為 10 且資料出現負偏時，也會出現類似最大估計誤差；當資料呈現常態時，其估計誤差最小。除了 G1 之外，其它的教育 Guni 指標均出現低估現象。G1 似乎比較適合於組別數小於 5 的情境。因此，指標的選擇與應用方式對於結論會產生重大的影響力。假如組別數等於或小於 5 或估計誤差無法容忍時，建議將 G3, G4, 與 G5 等係數乘以向上校正因子（n /(n-1)）。

***關鍵詞*：Gini 係數、Lorenz 曲線、教育成就、公平性測量**

[*] 本文第一作者（通訊作者）為國立嘉義大學教育學系教授
E-mail: fredli@mail.edu.tw

The educational attainment of society members is closely related to economic growth, social stability, social wellbeing, and people's health. As Harbison & Myers (1965) put it, "Education is both the seed and the flower of economic growth ". The topic of educational attainment divide, perhaps the most manifest evidence of income, earning, and opportunity inequality, has attracted much attention in recent years (Appiah-kubi, 2002; Checchi, 2001; Lopez, Thomas & Wang, 1998; Thomas, Wang & Fan, 2000 & 2002; Wahyuni, 2004) because equal access to education is widely viewed as an indispensable human right by which mobility in social classes can be facilitated so that social unrest or conflict can be lessened. Therefore, the educational divide in educational attainment is most concerned in developing and under-developed countries because it is "a critical underlying driver of the vicious circles of poverty observed at the household, regional, and national levels" (Perry, Arias, Lopez, Maloney & Serven, 2006).

To assess the magnitude of existing divide in educational attainment, various forms of a given Gini index have been developed by researchers in social science. The existence of these various computational forms has created several problems. First, in estimating educational inequality, it is difficult for a researcher to pick an optimal index from among these measures. Second, little is known about the relative size of estimated errors that arise from each of these inequality measures. Third, although Yitzhaki (1998) and Deltas (2003) reported that the use of grouped data caused " a downward bias in estimates of inequality", it is not clear whether these measures are sensitive to the true population form of education distribution and how many groups should be used for optimally plotting the Lorenz curve by which the Gini coefficients are estimated. Obviously, it is critical to pin-point which index is under-estimated or which index is over-estimated to take an educated policy of education. Therefore, identification of optimal measure(s) of Gini indices under different forms of education distribution and different data grouping sizes deserves further investigation.

Method

Inequality Measures

Lorenz curves, and the Gini coefficient are often used in income equity analyses that solely concern the dispersion of income distribution, independent of the mean of that distribution. Take Figure 1 as an example of education attainment, the Lorenz curve connects the cumulative proportion of the population at each level of education attainment in the horizontal axis with the corresponding cumulative proportion of schooling in the vertical axis (The dark boxes on the Lorenz curve stand for a given year of schooling).

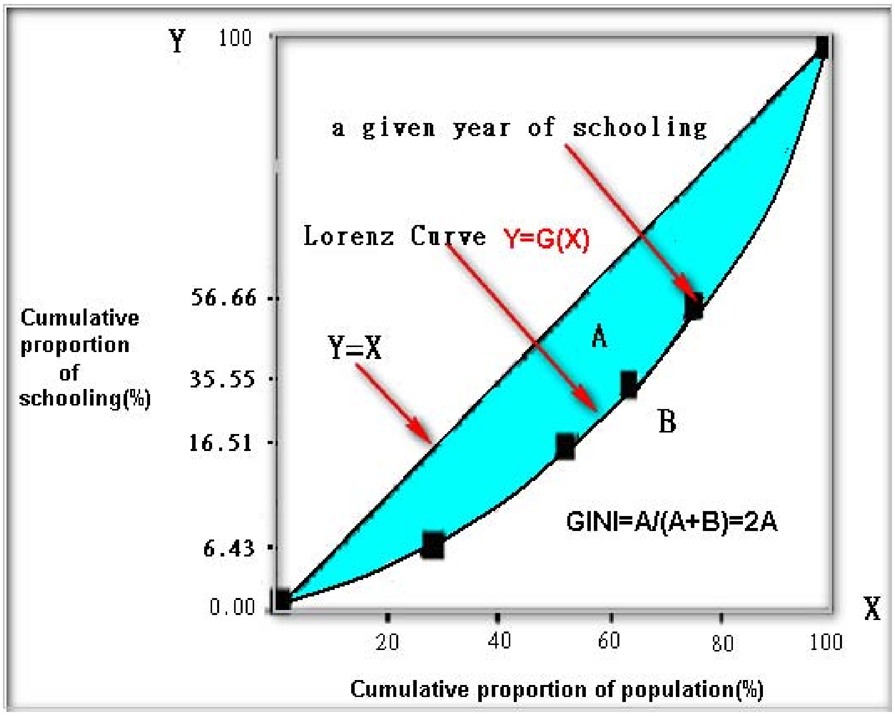

Figure 1. A Lorenz plot for schooling at each education level

Visually, the further the curve departs from the 45 degree line (indicating equal attainment), the greater the extent of inequality. The area under the Lorenz curve (B in Figure 1) can be obtained via definite integration if the exact mathematical function (LCDF in a linear, or quadratic, or cubic, higher-order form) representing the Lorenz curve exists. The area in the X-Y box is defined as 1.0 so that the area under the line Y=X is .5(i.e., A+B). Then, the Gini coefficient can be indirectly calculated as.

$$Gini = 2\int_0^1 (x - f(x))dx = 1 - 2\int_0^1 L_{CDF} = \frac{A}{A+B} = 2A$$

Thus, the geometric interpretation of the Gini coefficient is the Lorenz curve (Deaton, 1997; Xu, 2004).

The Gini Family

The Gini coefficient, originally developed by the Italian statistician, Gini (1912), is in essence a measure of variance. The extent of inequality displayed by the Lorenz curve (See Figure 1) can be measured by the Gini coefficient, which is defined as the ratio of the area between the Lorenz curve and the 45 degree line (A) over the total area under the 45 degree line (A+B), numerically equal to 2A (Xu, 2004). The Gini coefficient takes values between 0 and 1, with higher values indicating greater inequality. Generally speaking, a value of zero indicates complete equality； below 0.2: high equality；between 0.2~0.3：moderate equality; between 0.3~0.4：bearable；between 0.4~0.6：moderate inequality; above 0.6：high inequality; a value of 1:complete inequality. Since the Gini index in a developed country falls usually between .24 and .36. Thus, 0.4 is often regarded as an alerting cordon. As the Gini coefficient increases above .6, social unrest or violence may occur, because poverty or inequality tends to force people to fight for power or wealth (Kluge, 2001).

Because the exact numerical function for the Lorenz curve is not available, numerical definite integration of the area under the Lorenz curve is nearly impossible. Yet, a dozen alternative ways of computing Gini coefficients exist (Yitzhaki, 1998; Coulter, 1989). Xu (2004) reviewed 80 years of research on Gini's index and found that it can be computed in four basic forms: (1) geographic form, (2) mean difference form, (3) covariance form, and (4) matrix form. Further, within each of these four forms there are different ways to compute a Gini-based index. In the present study we focus on eleven well-known Gini-based indices that were investigated in Xu's (2004) literature review. They are classified in terms of data unit of analysis and listed below and renamed (inside parenthesis) for quick cross-reference:

I. Full sample raw data approaches

1. Jackknife mean Gini (M1)
2. pair-wised Gini's relative mean-difference without repetition (M2)

3. pair-wised Gini’s relative mean-difference with repetition (M3)

II. Grouped data approaches

1. Cumulative percentage/ratio differences (G1, G4)
2. Absolute values of relative ratio differences (G2)
3. Trapezoidal approximation approach (G3, G5),
4. Parabolic approximation approach (i.e., Simpson’s method, G6),
5. Absolute values of mean differences approach (i.e., Education Gini, G7).
6. Definite integration(DI)

In terms of data-entry methods, indices are calculated either from (1) unclassified individual raw data, as in M1, M2, M3 or (2) grouped data (usually in quintile or decile data, or in natural categories), as in G1~G7. With the grouped data, grouping or sorting of observations is required. They are usually in the form of relative or cumulative percentages. Due to the grouping of observations, differences within group are ignored (Lerman & Yitzhaki, 1989) so that grouped-indices will be under-estimates (downward biased). These 11 investigated indices of inequality are briefly defined below:

1. M1, also named as Jackknife mean Gini, is defined by Ogwang (2000)

Ogwang’s (2000) method for computing the Gini index and its modified Jackknife standard error entails the following steps：

(1) Get the target variable defined, such as income (INC),where the target variable is arranged in ascending order and create a dummy variable N representing a vector of 1,2,3,…,n,

(2) Generate the cumulative INC, such as CINC and compute a weighted INC , WINC=INC*N,

(3) Compute necessary summary statistics: total income, TINC and total weighted income, TWINC,

(4) Calculate the Gini index with all the N observations:
GNO=(2/N)×(TWINC/TINC)-1-1/N

(5) Calculate the Gini index with the kth observation deleted (Jackknife’s estimate) by the formula shown below:

$$G_{NK}(N,K)=G_{NO}+\frac{2}{N}\times\frac{INC[K]\times\frac{TWINC}{TINC}+\frac{TWINC}{N-1}}{TINC-INC[K]}-\frac{2}{N-1}\times\frac{TINC-CINC[K]+K\times INC[K]}{TINC-INC[K]}-\frac{1}{N(N-1)}$$

where K*INC[K]=WINC, CINC[K]= $\sum_{i=1}^{k} INC\ [i]$,

(6) Calculate the mean of the Jackknife's Gini indexes,M1=ΣGNK/N

(7) Compute the Jackknife standard error:

$$G_{JSE} = \sqrt{\frac{N-1}{N} * \sum_{k=1}^{n} (G_{NK}[N,K] - M1)^2}$$

For the normal distribution, as N increases, the standard error of Gini's mean difference can be also approximated by the following formula (Nair, 1936; Cowell, 1995).

$$G_{error} = \frac{.8068\, CV}{\sqrt{N}}$$

where CV is the coefficient of variation and is defined as:

$$CV = \frac{\sqrt{\frac{\sum_{i=1}^{N} (Y_i - \overline{Y})^2}{N}}}{\overline{Y}}$$

where N is the number of observations, Y_i is the i_{th} observation.

The standard error formula (G_{error}) can be applied to M2, M3, G7, and other types of indices that are derived from the individual observations without grouping.

2. M2, based on Gini's relative mean differences of pair-wise comparisons without repetition

Gini (1912) showed that the Gini index is closely related to Gini's relative mean difference. In fact, the Gini index can be expressed as half the Gini's relative mean difference. Due to the pair-wise comparisons without repetition of each data point itself, there are merely n(n-1)/2 pairs of comparisons (Stuart & Ord, 1987; Deaton, 1997; Xu, 2004). In the equation of M2 shown below, the absolute mean differences divided by the mean of the Y distribution is what Gini called the relative mean difference (in a way similar to the relative average deviation, which takes into account the mean for controlling central tendency). M2, also known as Gini's coefficient of concentration, is expressed as:

$$M2 = \frac{1}{\overline{Y} \times n(n-1)} \sum_{i>j} \sum_{j} \left| Y_i - Y_j \right| = \frac{1}{\overline{Y} \times 2} \frac{\sum_{i>j} \sum_{j} \left| Y_i - Y_j \right|}{\frac{n \times (n-1)}{2}} = \frac{\sum_{i}^{n} (2i - n - 1) \times Y_i}{(n-1) \times \sum_{i}^{n} Y_i}$$

where n is the number of observations, Y_i is the i_{th} observation. M2 is the typical Gini coefficient and can be calculated via any of the three approaches shown above, of which the last approach requires a sorted Y_i in a descending order.

3. M3, based on Gini's relative mean differences of pair-wise comparisons with repetition

M3 is also based on Gini's (1912) mean difference concept; yet, it includes all possible pairs of comparisons. This is the reason why the total mean difference is divided by n^2 rather than n(n-1). There are two approaches to compute M3.

$$M3 = 1 + \frac{1}{n} - \frac{2}{n^2 \times \overline{Y}}\left(\sum_{i}^{n} (n - i + 1) \times Y_i\right) = \frac{1}{2 \times n^2 \times \overline{Y}} \sum_{i=1}^{n} \sum_{j=1}^{n} | Y_i - Y_j |$$

where n is the number of observations, Y_i is an individual observation. For the first approach, descending sorting is required for the individual observations and each observation is weighted by its rank.

$$G1 = \frac{\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i}{(n \times 100) - \sum_{i=1}^{n} X_i}$$

4. G1, in cumulative percentages at each group of equal size, is expressed as where X_i & Y_i are cumulative percentages for X and Y, respectively, n is the number of the classified intervals. G1 as defined in the present study, has been empirically proved to be an unbiased estimate for small-samples (see later explanation in the results section).

5. G2, basically a summary measure of pair-wised deviation of data (Burt & Barber, 1996), is defined as

$$G2 = 0.5 * \sum_{i=1}^{n} \left| X_i - Y_i \right|$$

where X_i & Y_i are the relative ratio at the i_{th} category of X & Y, respectively. It can be viewed as the sum of a series of triangles and rectangles.

6. G3, estimated by the trapezoidal approximation to the integral, is expressed as

$$G3 = \left| 1 - \sum_{i=0}^{n-1} (X_{i+1} - X_i)(Y_{i+1} + Y_i) \right|$$

where X_i & Y_i represent the cumulative percentages of X & Y. G3 can be redefined as：

$$G3 = 2 \times \left| 0.5 - 0.5 * \sum_{i=0}^{n-1} (X_{i+1} - X_i)(Y_{i+1} + Y_i) \right|$$

where $X_0=0$，$Y_0=0$。

7. G4 is expressed as

$$G4 = \sum_{i=1}^{n}[2\times(X_i - Y_i)(X_i - X_{i-1})]$$

where $X_i=1/n$ (n is the number of intervals with equal size), Data on Y_i are the cumulative percentages for Y. Sorting is required for G4 (Left Business Observer, 1996).

8. G5, a trapezoidal estimate for the integral, is defined as

$$G5 =| 1-2\times(\frac{1}{n}\times\frac{1}{2}G(X_0)+G(X_1)+G(X_2)+\cdots+G(X_{n-1})+\frac{1}{2}G(X_n))|$$

The error in approximating an integral by the trapezoidal method will result in a value less than $\frac{|k|}{4n^2}$, k is the maximum value of the 1st derivative for G(x) (Espericueta, 2001). As the number of the intervals (n) increases, the estimation error decreases.

9. G6, the parabolic approximation approach, also named as Simpson's method, is defined by

$$G6 = 1-2(\sum_{k=1}^{n}(f_{(X_{K-1})}+4f(\frac{X_{K-1}+X_K}{2})+f_{(X_K)})(\frac{1}{6n}))$$

$$= 1-2(\sum_{k=1}^{n}(\frac{f_{(X_{K-1})}+4f(\frac{X_{K-1}+X_K}{2})+f_{(X_K)}}{6})(\frac{1}{n}))$$

$$= 1-2(\sum_{k=1}^{n}(\frac{f_{(X_{K-1})}}{6}+\frac{4}{6}f(\frac{X_{K-1}+X_K}{2})+\frac{f_{(X_K)}}{6})(\frac{1}{n}))$$

$$= 1-2(\sum_{k=1}^{n}(\frac{f_{(X_{K-1})}}{6}+\frac{4}{6}f(\frac{X_{K-1}+X_K}{2})+\frac{f_{(X_K)}}{6})(\frac{1}{n})$$

$$\cong 1-2(\sum_{k=1}^{n}(\frac{Y_{K-1}}{6}+\frac{4}{6}(\frac{Y_{K-1}+Y_K}{2})+\frac{Y_K}{6})(\frac{1}{n})))$$

$$= 1-2(\sum_{k=1}^{n}(\frac{Y_{K-1}}{6}+\frac{2}{6}(Y_{K-1}+Y_K)+\frac{Y_K}{6})(\frac{1}{n})))$$

$$= 1-2(\sum_{k=1}^{n}(\frac{1}{2}(Y_{K-1}+Y_K))(\frac{1}{n})))$$

Note the last term that reveals a close relationship with the Trapezoid approach (cf, G3 & G5). G6 can also be estimated by

$$G6 = 1-2(\sum_{k=1}^{n/2}\frac{(Y_{2K-2}+4Y_{2K-1}+Y_{2K})}{3}(\frac{1}{n}))$$

Since the exact parabolic function, $f((X_{k-1}+X_k)/2)$ cannot be known in advance, it is approximated by the average of $Y_{k-1}+Y_k$ as shown in the equation. This will lead to an identical result with the Trapezoid estimate as defined in the present study.

10. G7, Education Gini Coefficient

The conventional Gini index cannot be computed for education data, because (1) individual education attainment data points are likely not available, and (2) education attainment in years of schooling is a discrete variable (usually between 0~22 years) so that the continuous approximation of a truncated Lorenz curve (with a kinked line) has been deemed unnecessary and inappropriate. Therefore, Thomas, Wang, and Fan (2000 & 2002) developed a different formula to accommodate the special features of the schooling distributions. They called this newly-formulated index the Education Gini Index. The computation of M2 entails the following procedures: First, the years of schooling are classified into three cycles of education: primary, secondary, and tertiary. Then, each collected data point is assigned a specific monotonically increasing year of schooling value according to the following scheme:

1. illiterate(Y_1=0)
2. partial-primary($Y_2=0.5C_p$)
3. complete primary($Y_3=C_p$)
4. partial-secondary($Y_4= C_p + 0.5C_s$)
5. complete secondary($Y_5=C_p + C_s$)
6. partial-tertiary($Y_6=C_p + 0.5C_t$)
7. complete tertiary($Y_7=C_p + C_s+ C_t$)

where C_p is the cycle of the primary education (usually $C_p = 6$); C_s is the cycle of the secondary education (usually $C_s = 6$); C_t is the cycle of the tertiary education (usually $C_t = 4$).

To be consistent with the more common education system used world-wide, four cycles of schooling were adopted. The additional fourth cycle of education named “advanced” (C_a) is added for graduate schools. The years of schooling at this advanced level could extend from 2 to 6. Adopting this classification scheme sets a ceiling value of 22 on education attainment in the simulations which follow.

To be consistent with the simulation study design, 5 levels and 10 levels of schooling (between 0~22) are adopted (rather than the educational levels just described above) for use in the present study. Take 10 levels for example, the average year of education attainment and its standard deviation are defined by:

$$\overline{Y} = \sum_{i=1}^{n=10} P_i \overline{Y}_i$$

$$\sigma = \sqrt{\sum_{i=1}^{n=10} P_i (\overline{Y}_i - \overline{Y})^2}$$

$$G7 = \frac{N}{N-1} \frac{1}{\overline{Y}} \sum_{i=2}^{10} \sum_{j=1}^{i-1} P_i \times |\overline{Y}_i - \overline{Y}_j| \times P_j$$

where N is the total observations, P_i stands for the proportion of population with a given level of schooling. Note that $\overline{Y}_i$, $\overline{Y}_j$ are the average years of schooling at different educational attainment levels (not the number of years at different education levels as defined by Thomas, Wang, and Fan (2000), n is the number of categories/levels in education attainment data.

11. DI, definite integration

It is actually integrated using a theoretical Lorenz curve (See details below).

Simulation Design

Procedures for the data simulation

The data were simulated in the following steps:

(1) set up three types of education attainment distributions (positively skew, negatively skew and normal) for the Y axis,

(2) determine the theoretical distribution form for the X axis,

(3) pick the number of intervals, and

(4) estimate the optimal function G(x)(0≤G(x)≤1).

To compare the theoretical area under the Lorenz curve and its estimated area, an optimal integral function, G(x), derived from the Lorenz curve is needed. Through the area comparisons, different inequality measures estimated from the Lorenz curves can be evaluated.

Specifically, G(x), is estimated using the following steps:

(1) The uniform distribution of X is equally divided into 5 or 10 intervals (simulate horizontal equality). Get the cumulated percentage for each interval. This is the theoretical distribution.

(2) According to the pre-specified distribution of Y (positively skew, negatively skew and normal), divide it into 5 or 10 grouping intervals. Get the cumulated percentage for each interval. This is the observed distribution.

(3) Estimate the optimal G(x) by regressing Y, as criterion values, over X , as predicting values. The optimal G(X) function is selected to best fit the simulated data (R^2 is as large as possible). The line of best fit can be obtained using SPSS Non-linear Regression subroutine (pick the quadratic and cubic form , the ANOVA table term and constant not-included term).

Experimental Conditions

Three factors were manipulated: education attainment (low, average, high), dispersion of education attainment (large, moderate, small), and number of groups used (5 and 10). Mainly for the horizontal equality issue, uniform distribution is assumed for the X variable. With 5 or 10 groups, each group has a value of .2 or .1, respectively for the X variable. For the Y variable, a positively skewed distribution of education attainment was simulated to reflect under-developed countries, a normal distribution for developing countries, and a negatively skewed distribution for developed countries. Based on the study design, there are 3 x 3 x 2=18 types of data conditions in the present study, as shown in Table 1.

Table 1

Study design for the simulations of 18 conditions

Factors involved							
# of intervals(X axis)			*5*			*10*	
Dispersion of education(Y axis)		*L*[*]	*M*	*S*	*L*	*M*	*S*
	Positive (9)	1	2	3	4	5	6
Distribution	Normal (11)	7	8	9	10	11	12
Shape[**]	Negative (13)	13	14	15	16	17	18

Note: *L is referred to Large standard deviation (SD=5), M for moderate standard deviation (SD=3.5), S for small standard deviation (SD=2). **The average years of the positively skewed, normal and negatively skewed distribution of education attainment are 9, 11, and 13, respectively.

Data Generation Engine

To generate realistic data, a method using 4-parameters proposed by Ramberg, Tadikamalla, Dudewicz and Mykytka (1979) was adopted. It is defined by the percentile function shown below:

$$E = \lambda_1 + \frac{[p^{\lambda_3} - (1-p)^{\lambda_4}]}{\lambda_2}\,(0 \le p \le 1)$$

where p is a uniform random variable, λ_1 is a location parameter, λ_2 is a scale parameter, λ_3、λ_4 are shape parameters.

Several lambda parameters, as shown in Table 2, were chosen in accordance with typical types of distribution of education attainment specified in Table 1.

Table 2

Parameters set up for the three types of distributions in the study

	g1	g2	λ_1	λ_2	λ_3	λ_4
Distribution	Skewess	Kurtosis	Location	Scale	Shape	Shape
Normal	0	0	0	.1974	.1349	.1349
Positive	1.0	3.0	-.379	-.0562	-.0187	-.0388
Negative	-1.0	3.0	.379	-.0562	-.0388	-.0187

To generate data that mimic real education attainment data for most of the countries in the world, a linear transformation was performed on each generated value (E) and the re-scaled value (Y) was trimmed to make sure each data point fell between 0 and 22.

Y=SD x E + Mean

where SD and Mean are the standard deviation and average schooling, which are as specified in Table 1. The forgoing procedure steps were repeated 1000 times for each simulated condition.

To generate data that looks like data from underdeveloped countries (positively skewed), data from developing countries (normal distribution), and data from developed countries (negatively skewed), the three hypothetical levels of average schooling as 9, 11, and 13 years of education attainment were employed. This average schooling factor was crossed with three types of schooling dispersion, as 2, 3.5, and 5 years of standard deviation to reflect a small, a medium, and a large variance, respectively. In total, 18 data set conditions were generated in the study. Because the years of schooling is usually limited to a range of 0~22, any generated data point for education attainment beyond 22 was deleted. This truncation produced a slight deviation from pre-specified mean, variance, skewness, and kurtosis and a downward estimate in schooling average. As a result, the data sets used in the study were chosen usually with several runs of the program to pick data sets that fit the pre-specified parameters.

The theoretical model of a Lorenz curve and its definite integration for generated data of education attainment

As data is generated for a given condition and transformed into equal-size sub-group proportions, the optimal mathematical function, G(x), was estimated by non-linear regression. Take the 18th condition for example, given a uniform X variable classified into 10 intervals and with a standard deviation of 2 (small SD) and a mean of 13 (negative skewness), specific steps for obtaining the theoretical model of a Lorenz curve and its definite integration follow:

1. Computing descriptive statistics (as shown in Table 3) for a specific generation condition

Table 3

Descriptive statistics for the 18th study condition in Table 1

N	Mean	SD	Skewness	Kurtosis
1000	12.533	2.0187	-.81667	2.05957

Owing to the data truncated to range within 0~22, the summary statistics deviate slightly from the set-up condition (M=13, SD=2, Skewness=-1, Kurtosis=3).

2. Obtaining raw data points for each interval

The subtotal of education attainment within each interval is shown below:

849.00 1062.00 1157.00 1200.00 1254.00
1300.00 1318.00 1400.00 1441.00 1555.00

3. Getting the optimal function, G(x) and its integration

With the interval data transformed into cumulated proportions, Y, as a criterion and the cumulative proportions from the X (predicting) variable as shown in Table 5, the optimal function, $G(x)=.746713x+.254704x^2$, is obtained using SPSS non-linear regression. The area under the G(x) curve is integrated as ≈.45826. Its corresponding Gini coefficient can be calculated: 1- 2 x .45826≈.08348

$$\int_0^1 (.746713x + .254704x^2)dx = 0.746713\times\frac{1}{2} + .254704\times\frac{1}{3} \cong .45826$$

Based on the G(x) function, almost identical data set as shown in Table 4 can be duplicated (R^2=.99995) for a given condition.

Table 4

Cumulative percentages for data of negative distribution

X(CF%)	10	20	30	40	50	60	70	80	90	100
Y(F%)	6.773	8.471	9.230	9.572	10.003	10.37	5.137	11.168	11.495	12.404
Y(CF%)	.773	15.244	24.474	34.046	44.049	54.419	64.933	76.101	87.596	100.0

For space reasons, details about theoretical models (optimal G(X) function) and its definite integration for other simulated conditions are omitted here. Details are available from the first author upon request.

Criteria for index evaluation

Theoretically speaking, an index of inequality derived from the full sample data should reflect more accurately the inequality, both between-group and within-group. Accordingly, M1 (Jackknife mean Gini), M2, M3 should perform better than any index of inequality derived from grouped data, where within-group inequality is ignored. In addition, any index of inequality computed by definite integration should perform better than any index of inequality approximated by the trapezoidal, the parabolic, and the other remaining approaches. Therefore, definite integration (DI), M1, M2, and M3 were used as index evaluation criteria. For practical use in a small sample, M2 was adjusted upward by a small-sample component, 1/(n(n-1)), instead of $1/n^2$, as recommended by Deltas (2003). He proposed that the Gini coefficient should be adjusted by n/(n-1) in the small sample setting to reduce "small-sample downward bias". This principle was applied to the grouped-data settings as well.

The criteria used to evaluate the performance of the Gini indices under various simulation conditions are:

1. the difference between DI and M1, M2, M3,
2. the estimation error between DI and its approximated methods,
3. index critical features (etc., decomposability, available standard error, and sensitive to group size, average years of schooling, and schooling dispersion).

RESULTS

The results of the simulation study are broken down by the size of SD of education attainment simulated. Following reports of the indices themselves are maximum/minimum error reports for each SD broken down by the groupings size of 10 or 5.

Impact of Average Schooling Skew for Specific Dispersions

Tables 5, 6 & 7 report the calculated Gini summary statistics and show the impact of average schooling with SD = 2 for Table 5, SD = 3.5 for Table 6, and SD = 5 for Table 7. Within each table the average years of schooling for undeveloped, developing, and developed countries are given as 9, 11, and 13 years respectively.

Table 5

The Gini coefficients based on the various estimation approaches by grouping size, and schooling level（SD=2）

Schooling levels	(positive) 9 yrs		(normal) 11 yrs		(negative) 13 yrs	
full sample approach						
M2	.13018		.10801		.08662	
M3, M1*	.13005		.10790		.08654	
Grouping size	10	5	10	5	10	5
I. DI**	.12784	.12601	.10608	.10426	.08348	.08290
II. Gini family						
G1	.1417	.1515	.1182	.1261	.0941	.1009
G2	.0902	.0902	.0774	.0730	.0595	.0595
G3、G4、G5	.1275	.1212	.1064	.1008	.0847	.0808
G6、G7	.1275	.1212	.1064	.1008	.0847	.0808

* M1 is an alternative way of computing Jackknife mean Gini proposed by Ogwang (2000). ** indicates definite integration.

Table 6

The Gini coefficients based on the various estimation approaches by group size, and average schooling（SD=3.5）

Schooling levels	(positive) 9 yrs		(normal) 11 yrs		(negative) 13 yrs	
full sample approach						
M2	.25689		.19792		.15778	
M3、M1	.25664		.19773		.15762	
Grouping size	10	5	10	5	10	5
I. DI	.2530	.2508	.1947	.1936	.1470	.1469
I. DI**	.12784	.12601	.10608	.10426	.08348	.08290
II. Gini family						
G1	.2803	.3018	.2169	.2338	.1659	.1789
G2	.1783	.1783	.1401	.1359	.1061	.1061
G3、G4、G5	.2523	.2415	.1952	.1870	.1493	.1432
G6、G7	.2523	.2415	.1952	.1870	.1493	.1432

Table 7

The Gini coefficients based on the various estimation approaches by interval size, and average schooling（SD=5）

Schooling levels	(positive) 9 yrs		(normal) 11 yrs		(negative) 13 yrs	
full sample approach						
M2	.29091		.25384		.19030	
M3、M1	.29061		.25359		.19011	
Grouping size	10	5	10	5	10	5
I. DI ≈	.28726	.28498	.24934	.24837	.18270	18217
I. DI**	.12784	.12601	.10608	.10426	.08348	.08290
II. Gini Family						
G1	.3176	.3425	.2776	.3000	.2052	.2215
G2	.2045	.2045	.1801	.1742	.1306	.1300
G3、G4、G5	.2858	.2740	.2499	.2401	.1846	.1772
G6、G7	.2858	.2740	.2499	.2401	.1846	.1772

As shown in each of Tables 5, 6 and 7 there are several clear and consistent patterns in the data.

1. Any measure of inequality derived from the full sample data, under various distribution forms, almost produces an identical coefficient, as appeared in M1, M2, and M3.
2. G3, G4, G5, G6, and G7 across various forms of distribution and group sizes all resulted in an identical value. Following the definitions used in G6, index values for G3, G4, G5, G6 should be identical (See explanation in the end of G6's definition). The unexpected finding for G7 requires further study.
3. The standard error of estimation was smaller from the full sample approach, like in Jackknife mean Gini, than the grouped data approach, like in G7. This may be due to the fact that within-group inequality is ignored in grouped data.

Interestingly, if G3~G7 index estimates are multiplied by an upward factor, n/(n-1), each is numerically equivalent to G1 under each simulation condition (e.g., .1275x(10/9) = .1417 for the 9 year positive 10-group condition in Table 5). It indicates that G1 is an unbiased estimator for smaller grouping sizes. This unexpected finding also deserves further study. The upward adjustment factor, n/(n-1), suggested by Deltas(2003) , is usually recommended for a small-sample setting or when the number of groups used is limited.

Error Analysis for Grouped Data

Error rates were somewhat more distinctive regarding factors manipulated in the study. Using each of M1, M2, M3, and DI (definite integration) as a criterion, the error values between the specific criterion and the other evaluated indexes under the group sizes of 10 and 5 are further analyzed for each size of variance. In order to easily identify the different sources of error of estimation, the maximum error and minimum error for each index under each simulation condition are compared and summarized in Table 8. Specific error patterns are analyzed under 3 sizes of SD as follows.

Table 8

Combined conditions for maximum and minimum error of estimation produced by Gini-based indices

Index Error Range	Error Type	SD	Distribution	Group Size	Criteria	Absolute
G1/G2	Max	2	Positive	5	All*	.0213~.0400
	Max	3.5	Positive	5	All	.0449~.0786
	Max	5	Positive	5	All	.0516~.0864
	Max	2	Positive	10	All	.0115~.0400
	Max	3.5	Positive	10	All	.0234~.0786
	Max	5	Positive	10	All	.0264~.0864
	Min	2	Negative	5	All	.0143~.0271
	Min	3.5	Negative	5	All	.0211~.0517
	Min	5	Negative	5	All	.0312~.0603
	Min	2	Negative	10	All	.0075~.0271
	Min	3.5	Negative	10	All	.0081~.0517
	Min	5	Negative	10	All	.0149~.0597
G3/G4 G5/G6 /G7	Max	2	Positive	10	M1/M2/M3	.0025~.0027
	Max	2	Negative	10	DI	.0012
	Max	3.5	Negative	10	All	.0023~.0085
	Max	5	Negative	10	All	.0019~.0057
	Max	2	Positive	5	All	.0048~.0090

續後頁

接前頁

	Max	3.5	Positive	5	All	.0093~.0154
	Max	5	Positive	5	All	.0110~.0169
	Min	2	Normal	10	All	.0003~.0016
	Min	3.5	Normal	10	All	.0005~.0027
	Min	5	Normal	10	All	.0006~.0039
	Min	2	Negative	5	All	.0021~.0058
	Min	3.5	Normal	5	M1/M2/M3	.0107~.0109
	Min	3.5	Negative	5	DI	.0037
	Min	5	Negative	5	All	.0050~.0131

Note: * applicable to all criteria: M1、M2、M3、and DI.

SD=2.

In general, several common error patterns found in Table 8 are:

1. Gini family indices displayed the greatest error for the positively skewed distribution of education attainment. This implies that measures of inequality will be least accurate when the distribution of education attainment is positively skewed (a scenario for under-developed countries).

2. No matter whether the grouping size is 10 or 5, the difference between the definite integration estimate and the trapezoidal estimate (not reported here) is less than the expected value, $\frac{|k|}{4N^2}$ (Espericueta, 2001). Note that

$$\frac{k}{4n^2} = \frac{k(b-a)}{4n^2}$$

where k refers to the maximum value of the 1st derivative for the Lorenz curve ($K \geqq |G'(x)|$), n is the grouping size. In the present study, the range of integration is between 0 and 1. Thus, b=1, a=0. Since the Lorenz curve is always an increasing function, its maximum value will fall at x=1. For example, k can be obtained by computing the 1st derivative for the $G'(x)$=.746713 + 2*.254704x. Under this function, k=1.256121 if x=1. G(x) is the optimal function for a given simulation condition.

3. Except for the over-estimated G1, compared with the indices derived from the full

sample data, the Gini-based indices are all under-estimated for all conditions where grouped data are used. This may be due to the within-group inequality ignored in the data analysis (Lerman & Yitzhaki, 1989). This is consistent with Deltas's (2000) report that: "the use of grouped data caused a downward bias".

4. Data with a grouping size of 10 resulted in smaller estimated errors than data with a grouping size of 5. As expected, index accuracy increases with the number of grouping intervals.

SD=3.5

Again, considering M1, M2, M3, and DI as a criterion, the error scores between the criterion and the other evaluated indices under the group size of 10 and 5 are further analyzed and significant patterns are found below:

1. The Gini-based indices display the greatest error with the positively-skewed distribution of education attainment when the grouping size is 5; while they demonstrate maximum error with the negatively skewed data when the grouping size is 10. The extent of error of the Gini-based indices of inequality depends on both grouping size and generated data shape.
2. No matter whether the grouping size is 10 or 5, the difference between definite integration estimate and the trapezoidal estimate is less than the expected value, $\frac{|k|}{4N^2}$ (Espericueta, 2001).
3. Except for the over-estimated G1, compared with indices derived from full sample data, the Gini-based indices are all under-estimated across all conditions using grouped data. This may be due to the ignored within-group inequality (Lerman & Yitzhaki, 1989). Deltas (2000) also reports: "the use of grouped data caused a downward bias".
4. Data with a grouping size of 10 would result in smaller estimated errors than data with a grouping size of 5. Index accuracy increases as the number of grouping intervals increases.

SD=5

Use M1, M2, M3, and definite integration as a criterion, the error scores between the criterion and the other evaluated indexes under the group size of 5 and 10 are further analyzed below:

1. G1, and G2 indices display the greatest error with the positively skewed distribution of education attainment regardless of grouping size. This implies that the measures of inequality will be least accurate when the distribution of education attainment is positively skewed (here again, this is a scenario often found in under-developed countries). Yet, the Gini-based indices do not produce clear patterns of error. For a grouping size of 10, the maximum error occurs with the negatively skewed distribution, while for a grouping size of 5, it occurs for the positively skewed distribution. Thus, there is no clear minimum error consistently found on the Gini indices.
2. No matter the grouping size (5 or 10), the difference between definite integration estimate and the trapezoidal estimate is less than the expected value, $\frac{|k|}{4N^2}$ Espericueta, 2001).
3. Except for the over-estimated G1, compared with indices derived from the full sample data, Gini-based indices are all under-estimated across all conditions when based on grouped data. This may be due to the ignored within-group inequality (Lerman & Yitzhaki, 1989). Deltas's (2000) reported: "the use of grouped data caused a downward bias"..
4. Index estimation error increases with the SD of the educational attainment distribution as shown in Table 8 as SD increases from 2 to 3.5 to 5.
5. Again, data with a grouping size of 10 would result in smaller estimated errors than data with the grouping size of 5. That is, index accuracy also increases with the number of groups used.

As shown in Table 8, G1, and G2 demonstrate greatest estimation error with positively skewed data and least estimation error with negatively skewed data, regardless of variance, grouping size, and criterion index used. The other Gini-based indices G3, G4, G5, G6, G7 produce similar patterns of maximum error as in G1 and G2 when the grouping size is 5; yet, when the grouping size is 10, they tend to display greatest error with negatively skewed data and least error with normal data. These findings do not consistently agree with Figini's (1998) results (See Figini's Table 2). In inequality studies, we are more concerned about the maximum error condition rather than the minimum error condition. Therefore, we should pay more attention to the conditions that produce the maximum error for an index. Absolute error range (max~min) is also shown in Table 8 for each type of error. It can be used to decide whether the estimated error for an index is tolerable.

CONCLUSIONS

Because only a limited number of testing conditions were investigated, generalization of these findings may be restricted. However, several major conclusions can be drawn from the results:

1. Index effectiveness will be influenced in general by the number of groups, forms of distribution and size of variance involved. The accuracy of inequality measures increases as the number of groups increases or as the size of variance decreases. When the distribution of education attainment is positively skewed and the number of groups is 5, Gini indices will produce their maximum error of estimation. When the distribution of education attainment is negatively skewed, the number of groups is 10, and the SD= 3.5 or 5.0, most Gini indices (except G1 and G2) will produce their maximum error of estimation.
2. Among the Gini indices derived from grouped data , G3, G4, G5, G6, and G7 perform equally well in terms of maximum error of estimation.
3. Perhaps, due to the large sample used in the study (N=1000), the outcomes of M1, M2, M3 derived from the full sample data appear almost the same, especially for the pair of M1 and M3. M1 and M3 are the best options when the analyzed data are individual observations because their standard errors can be computed.
4. Except the over-estimated G1, the other Gini indices were all under-estimated. This downward bias of estimates might be attributed to ignored intra-group inequality (Lermann & Yitzhaki, 1989) and downward small-sample bias (Deltas, 2003). And their estimated errors are less than $\frac{|k|}{4N^2}$. This finding is in accord with Espericueta's theory (2001). Ourti and Clarke (2011) also recommended a correction term($n^2/(n^2-1)$, n=# of groups) to remove the main bias of the Gini index due to grouping.
5. Interestingly, the modified education Gini index, G7, appears the best option when data

are organized into groups. This is important since G7 is the only grouped-data index for which one can compute a standard error.

6. G1 remained an unbiased estimate of inequality for the smallest grouping size of 5.

These conclusions have important implications for investigating educational attainment inequality in practice. First, to reduce error of user-estimation, the number of groups employed should be more than 5. If 5 or fewer groups are used or the estimated error is not tolerable, it is recommended that one upwardly adjust the G3, G4 or G5 estimates by a factor of n/(n-1). Interestingly, with this upward adjustment, they were numerically equivalent to the G1 index. Accordingly, G1 is recommended whenever a smaller size of grouping is necessary. Third, interpretation of an index's under-estimation should be considered, especially dealing with a positively skewed distribution with large variance (a scenario often found in under-developed countries). Fourth, G7 is recommended for the grouped-data case, while M1, M2, M3 are recommended for data with individual observations.

References

外文部分

Appiah-kubi, K. (2002). *Education inequality in Ghana: Gini coefficient of education.* Paper presented at the MIMAP Meeting , Pavillon La Laurentienne, Universite Laval, Ste. Foy. Quebec.

Burt, J. E., & Barber, G. M. (1996). *Elementary statistics for Geographers.* New York: Guilford Press.

Checchi, D. (2001). *Education inequality and income inequality.* Retrieved July 6, 2005 from the World Wide Web: *http://sticerd.lse.ac.uk/publications/darp.asp*

Coulter, P. B. (1989). *Measuring inequality*. Boulder: Westview Press.

Cowell, F. A. (1995). *Measuring inequality(2nd ed.)*. Hemel Hempstead: Harvester Wheatsheat.

Deaton, A. (1997). *The analysis of household surveys: A microeconomic approach to development policy.* Johns Hopkins University Press, Baltimore and London.

Deltas, G. (2000). *The small sample bias of the Gini coefficient: Results and implications for* *empirical research.* University of Illinois, Urbana-Champain. Retrieved December 1, 2001 from the World Wide Web : *http://www.staff.uiuc.edu/~deltas.*

Deltas, G. (2003). The small sample bias of the Gini coefficient: Results and implications for empirical research. *The Review of Economics and Statistics, 85(1)*, 226-234.

Espericueta, R. (2001). *Numerical integration theorems*. Retrieved December 1, 2001 from the World Wide Web :
http://online.bc.cc.ca.us/mathb6b/content/chapter_notes/chapter7/theorems/summary.htm

Figini, P. (1998). *Measuring inequality: on the correlation between indices.* Trinity Economic Papers, Technical Paper No. 98/7, Trinity College Dublin.

Gini, C. (1912) *Variabilità e Mutabilità(Variability and Mutability): Contributo allo Studio delle distribuzioni e delle relazioni statistiche.* Facoltá di Giurisprudenza della R. Universitá dei Cagliari, anno III, parte 2.

Harbison, F., & Myers, C. (1965). *Manpower and education.* New York: McGraq-Hill.

Kluge, G. (2001). *Trickle down trash, squeeze up wealth.* Retrieved November 12, 2002 from the World Wide Web: *http://poorcity.richcity.org/entundp.htm*

Left Business Oberver (1996). *Gini says: Measuring income inequality.* Retrieved January 2, 2004 from the World Wide Web:
http://www.leftbusinessobserver.com/Gini_supplement.htm.

Lerman, R., & Yitzhaki, S.(1989). Improving the accuracy of estimates of Gini coefficient. *Journal of Economics, 42,* 43-47.

Li, M. N. (2004). A relative accuracy analysis of education inequality measures derived from the Gini family. *Journal of Research on Elementary and Secondary Education, 13*, 1-50.

Lopez, R., Thomas, V., & Wang, Y. (1998). *Addressing the education puzzle: The distribution of education and economic reform.* Policy research working paper No. 2031. Washington, D C: World Bank.

Nair, U. S. (1936). The standard error of Gini’s mean difference. *Biometrika, 28*, 428-436.

Ogwang, T. (2000). A convenient method of computing the Gini index and its standard error. *Oxford Bulletin of Economics and Statistics, 62(1*), 123-129.

Ourti, T. M., Clarke, P. (2011). A Simple Correction to Remove the Bias of the Gini Coefficient Due to Grouping. *The Review of Economics and Statistics, 93(3)*, 982-994.

Perry, G. E. Arias, O. S., López, J. H., Maloney, W. F., Servén, L. (2006). *Poverty reduction and growth: Virtuous and vicious circles.* Washington, DC, NW: The World Bank.

Ramberg, J. S., Tadikamalla, P. R., Dudewicz, E. J., & Mykytka, E. F. (1979). A probability distribution and its uses in fitting data. *Technometrics, 21(2)* , 201-214.

Stuart, A., & Ord, K. J. (1987). *Kendall's advanced theory of statistics*(5th ed.). New York: Oxford University Press.

Thomas, V., Wang, Y., & Fan, X. (2000). *Measuring education inequality: Gini coefficients of education.* Washington, DC: World Bank Institute, World Bank.

Thomas, V., Wang, Y., & Fan, X. (2002). *A new dataset on inequality in education: Gini and theil indices of schooling for 140 countries, 1960-2000*. Retrieved July 1, 2003 from the World Wide Web: *http://www33.brinkster.com/yanwang2/EduGini-revised10-25-02.pdf*

Wahyuni, A. S. (2004). *Education inequality in Indonesia 1980-2000.* Paper presented at the 6th IRSA international conference. Yogyakarta, Indonesia.

Xu, K. (2004). *How has the literature on Gini's index evolved in the past 80 years?* Retrieved Nov 10, 2004 from the World Wide Web: *http://economics.dal.ca/RePEc/dal/wparch/howgini.pdf*

Yitzhaki, S. (1998). More than a dozen alternative ways of spelling Gini. *Research on Economic Inequality, 8,* 13-30.

Note: 1. Special thanks go to Dr. Gary Lautenschlager for his valuable advice on research design.

2. This paper is an adapted English version of my original Chinese article(Li, 2004).

文稿收件：2012 年 09 月 06 日
文稿修改：2012 年 11 月 19 日
接受刊登：2013 年 01 月 25 日

Geni-Derived Index Effectiveness in Educational Inequality Measurement

Mao-Neng Li[*]

Abstract

In a simulation study, eleven frequently-used Gini indices of educational inequality were investigated for their sensitivity to educational attainment distribution and the number of groups used for data analysis.In general, index effectiveness was influenced by the number of grouping, forms of distribution, and size of variance involved. The accuracy of inequality measures increases as the number of groups increases or as the size of variance decreases. G1, and G2 demonstrate greatest estimation error with positively skewed data and least estimation error with negatively skewed data, regardless of variance, grouping size, and criterion index used. The other Gini-based indices G3, G4, G5, G6, G7 produce similar patterns of maximum error with postively skewed data when the grouping size is 5; yet, when the grouping size is 10, they tend to display greatest error with negatively skewed data and least error with normal data. Indices of Gini are all under-estimated except for the G1. G1 seems an unbiased index for a small grouping size less than 5. Therefore, the choice of index and method of implementation can have critical bearing on the conclusions reached. If 5 or fewer groups are used or the estimated error is not tolerable, it is recommended that one upwardly adjust the G3, G4 or G5 estimates by a factor of n/(n-1).

***Keywords*: Gini Coefficient, Lorenz Curve, Education Attainment, Inequality Measurement**

[*] Professor, Department of Education, National Chiayi University.
E-mail: fredli@mail.ncyu.deu.tw