An Investigation of the Impact of Guessing on Coefficient α and Reliability

Applied Psychological Measurement
2015, Vol. 39(4) 264–277
© The Author(s) 2014
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621614559516
apm.sagepub.com

Insu Paek¹

Abstract

Guessing is known to influence the test reliability of multiple-choice tests. Although there are many studies that have examined the impact of guessing, they used rather restrictive assumptions (e.g., parallel test assumptions, homogeneous inter-item correlations, homogeneous item difficulty, and homogeneous guessing levels across items) to evaluate the relation between guessing and test reliability. Based on the item response theory (IRT) framework, this study investigated the extent of the impact of guessing on reliability under more realistic conditions where item difficulty, item discrimination, and guessing levels actually vary across items with three different test lengths (TL). By accommodating multiple item characteristics simultaneously, this study also focused on examining interaction effects between guessing and other variables entered in the simulation to be more realistic. The simulation of the more realistic conditions and calculations of reliability and classical test theory (CTT) item statistics were facilitated by expressing CTT item statistics, coefficient α , and reliability in terms of IRT model parameters. In addition to the general negative impact of guessing on reliability, results showed interaction effects between TL and guessing and between guessing and test difficulty.

Keywords

guessing, reliability, coefficient alpha, item response theory

When investigating psychometric properties of a test with dichotomously scored items, coefficient α or Cronbach α (Cronbach, 1951) is often used to summarize one of the most important psychometric properties of a test as a reliability index (precisely speaking reliability for essentially tau-equivalent items in the test). For a multiple-choice test with dichotomous scoring, as is popular in many educational achievement tests, guessing may influence item responses. The popular coefficient α —and classical test theory (CTT) item statistics such as item easiness and item-total correlation (ITC) or item-rest correlation (IRC), where the rest is the total test score excluding the score of the item of interest—is widely used for multiple-choice tests, but it is not designed to accommodate guessing. Potentially, guessing could bring significant changes in coefficient α and the CTT item statistics, affecting the overall test quality. Given that coefficient α is routinely calculated as a reliability measure in practice for multiple-choice tests, it

¹Florida State University, Tallahassee, USA

Corresponding Author:

Insu Paek, Educational Psychology and Learning System, Florida State University, 3204D Stone Building, 1114 W. Call St., Tallahassee, FL 32306-4453, USA.

Email: ipaek@fsu.edu

would be interesting and informative to understand how much impact guessing has on coefficient α and reliability in general. In the past, researchers have studied the role of guessing on multiple-choice tests (e.g., Angoff, 1989; Carroll, 1945; Gross & Wright, 1985; Kim & Feldt, 2010; Lord, 1944, 1952; MacCann, 2004; Plumlee, 1952; Zimmerman & Williams, 2003), and mostly they focused on the impact of guessing on reliability and the inter-item correlations, relating the number of options or test length (TL) to test reliability. All the studies above (except Kim & Feldt's, 2010, study) are fundamentally based on CTT. In addition, all the studies above used rather restrictive assumptions (e.g., parallel test assumptions, homogeneous inter-item correlations, homogeneous item difficulty, and homogeneous guessing levels across items) to derive or show the relation between guessing and reliability. The use of the item response theory (IRT) framework to assess the role of guessing in multiple-choice tests for reliability is quite rare in the literature. In this study, based on the IRT framework, the author investigated the extent of the impact of guessing on coefficient α and reliability under more realistic conditions where item difficulty, item discrimination, and guessing levels actually vary across items with three different test length (TL).¹ Because the unit of modeling in IRT is at the item level, IRT provides a scaffold to express marginal observed item and test score statistics (e.g., proportion correct, ITC or IRC, mean test score), coefficient α , and reliability as a function of the IRT model parameters, especially as a function of guessing level. In addition, by employing the IRT framework, this study does not need to assume aforementioned restrictive assumptions used in the previous studies. This study employed the three-parameter logistic model (3PLM; Birnbaum, 1968) for the item response function (IRF), which describes the probability of correct response given person ability.

The major theme of this study centers on examining the degree of the impact of guessing on coefficient α and reliability for a variety of tests having different test properties. To this end, many tests having different difficulties, discriminations, and guessing levels are constructed based on certain populations of those IRT item and person parameters where tests' properties are drawn. Analyses and inferences are made at the population level, that is, investigations of expected behaviors of coefficient α and reliability under the effect of guessing at the population level. Coefficient α and reliability were differentiated because the use of the IRT framework where items have different discrimination, difficulty, and guessing levels does not lead to the essentially tau-equivalent items (or even the congeneric items). Also, it is because the concept of reliability, which is the ratio of the true score variance to the observed score variance, can be directly defined by the IRT model parameters. The computation of CTT item statistics, coefficient α , and reliability was conducted by direct evaluation of the expressions of these statistics, which are function of the IRT model parameters. This approach permits efficient computing with easy implementation of a variety of conditions at the population level because actual item responses do not need to be generated.

Item, Test Statistics, and Reliability

Across all formulas and derivations, the local independence was assumed as usual in IRT, and the person ability distribution $g(\theta)$ was assumed to be the standard normal distribution. Let $U_i = i$ th item response (0 for incorrect or 1 for correct) and $X = \sum_i U_i$.

CTT Item Difficulty and Item Discrimination

The 3PLM is

$$E(U_i|\theta) = P(U_i = 1|\theta) = P_i(\theta) = P_i = g_i + (1 - g_i) \frac{\exp[Da_i(\theta - b_i)]}{1 + \exp[Da_i(\theta - b_i)]}, \quad (1)$$

where a_i is the i th item slope, b_i is the i th item location parameter as item difficulty, g_i is guessing for the i th item, and D is a scaling constant equal to 1.7.

$$E(U_i) = \pi_i = \int P_i(\theta)g(\theta)d\theta. \quad (2)$$

This π_i serves as CTT item difficulty (item easiness). For ITC as an item discrimination, using the covariance decomposition formula

$$ITC_i = \frac{\text{Cov}(X, U_i)}{\sqrt{\text{var}(X)\text{var}(U_i)}} = \frac{\text{Cov}[E(X|\theta), E(U_i|\theta)] + E[\text{Cov}(X, U_i|\theta)]}{\sqrt{\text{var}(X)\text{var}(U_i)}}. \quad (3)$$

$$\begin{aligned} \text{Cov}[E(X|\theta), E(U_i|\theta)] &= E\{[E(X|\theta) - E(E(X|\theta))][E(U_i|\theta) - E(E(U_i|\theta))]\} \\ &= E\{[E(X|\theta) - E(X)][E(U_i|\theta) - E(U_i)]\} \\ &= E\left[\left(\sum P_i(\theta) - \mu_X\right)(P_i(\theta) - \pi_i)\right] \\ &= \int \left(\sum P_i(\theta) - \mu_X\right)(P_i(\theta) - \pi_i)g(\theta)d\theta, \text{ and} \end{aligned} \quad (4)$$

$$\begin{aligned} E[\text{Cov}(X, U_i|\theta)] &= E\left[\text{Cov}\left(\sum_j U_j, U_i|\theta\right)\right] \\ &= E\left[\text{Cov}(U_i, U_i|\theta) + \sum_{i \neq j} \text{Cov}(U_j, U_i|\theta)\right] \\ &= E[\text{Cov}(U_i, U_i|\theta)] \\ &= E[\text{var}(U_i|\theta)] \\ &= \int P_i(\theta)Q_i(\theta)g(\theta)d\theta, \end{aligned} \quad (5)$$

where $Q_i(\theta) = 1 - P_i(\theta)$. The second and the third lines of Equation 5 are due to the local independence assumption. After defining $\text{var}(U_i) = \sigma_i^2$, μ_X , and $\text{var}(X) = \sigma_X^2$ in terms of the $P_i(\theta)$ and $Q_i(\theta)$, plugging Equations 4 and 5 into Equation 3,

$$\begin{aligned} ITC_i &= \left(\int \sum_i P_i(\theta)Q_i(\theta)g(\theta)d\theta + \int \left[\sum_i P_i(\theta) \right]^2 g(\theta)d\theta - \left[\int \sum_i P_i(\theta)g(\theta)d\theta \right]^2 \right)^{-1/2} \\ &\times \left(\int P_i(\theta)Q_i(\theta)g(\theta)d\theta + \int [P_i(\theta)]^2 g(\theta)d\theta - \left[\int P_i(\theta)g(\theta)d\theta \right]^2 \right)^{-1/2} \\ &\times \left(\int \left(\sum P_i(\theta) - \mu_X \right) (P_i(\theta) - \pi_i)g(\theta)d\theta + \int P_i(\theta)Q_i(\theta)g(\theta)d\theta \right). \end{aligned} \quad (6)$$

IRC is obtained when we define $X \equiv X_{(-i)}$, where $X_{(-i)}$ is the sum of all item responses except the i th item in Equations 3 through 6. Note that when $X_{(-i)}$ replaces all X , $E[\text{Cov}(X_{(-i)}, U_i|\theta)]$ is zero and $\int P_i(\theta)Q_i(\theta)g(\theta)d\theta$ in the numerator disappears. In this study, IRC was used as a CTT item discrimination index.

For π_i , its relationship with guessing (g) will be linear given a_i and b_i . When π_i is approximated by a numerical integration with the quadrature point θ_k^* covering practically reasonable θ range

$$\pi_i = \int P_i(\theta)g(\theta)d\theta \approx \sum_{\theta_k^*} P_i(\theta_k^*)w(\theta_k^*) = \sum_{\theta_k^*} P_i^*(\theta_k^*)w(\theta_k^*) + g_i \sum_{\theta_k^*} Q_i^*(\theta_k^*)w(\theta_k^*), \quad (7)$$

where $P_i^* = (1 + \exp[-Da_i(\theta - b_i)])^{-1}$, $Q_i^* = 1 - P_i^*$, $w(\theta_k^*)$ is a weight at θ_k^* , and $\sum_{\theta_k^*} w(\theta_k^*) = 1$.

When there is no guessing ($g_i = 0$), π_i will be the value from the intercept term in the above equation. It is clear that π_i would increase linearly at the rate of the slope in the above equation.

Test Reliability

In CTT, $X = \tau$ (true score) + ε (error). Because the variance of error of measurement (VEM), which is the square of the standard error of measurement (SEM), is $VEM = SEM^2 = \sigma_\varepsilon^2 = \int \sum_i P_i(\theta)Q_i(\theta)g(\theta)d\theta$, the IRT reliability for X is

$$\begin{aligned} \text{IRT.rel.X} &= 1 - \sigma_\varepsilon^2 / \sigma_X^2 \\ &= 1 - \frac{\int \sum_i P_i(\theta)Q_i(\theta)g(\theta)d\theta}{\int \sum_i P_i(\theta)Q_i(\theta)g(\theta)d\theta + \int \left[\sum_i P_i(\theta) \right]^2 g(\theta)d\theta - \left[\int \sum_i P_i(\theta)g(\theta)d\theta \right]^2} \\ &= \frac{\int \left[\sum_i P_i(\theta) \right]^2 g(\theta)d\theta - \left[\int \sum_i P_i(\theta)g(\theta)d\theta \right]^2}{\int \sum_i P_i(\theta)Q_i(\theta)g(\theta)d\theta + \int \left[\sum_i P_i(\theta) \right]^2 g(\theta)d\theta - \left[\int \sum_i P_i(\theta)g(\theta)d\theta \right]^2}. \end{aligned} \quad (8)$$

Lord (1977, 1980) also showed how to estimate the test reliability for number-right scores with given IRT item parameters and individual ability parameters. The above reliability formula can be seen as an adaptation to a population case where θ is from a distribution $g(\theta)$. Note that the IRT reliability can be defined also for $\hat{\theta}$ (see, e.g., Green, Bock, Humphreys, Linn, & Reckase, 1984; Nicewander & Thomasson, 1999; Samejima, 1994), but the present article restricts its interest to the reliability for the observed score X .

Finally, the well-known coefficient α (C_α) is $[N/(N-1)] \times [1 - (\sum \sigma_i^2)/\sigma_X^2]$, where N is the number of items in a test and σ_i^2 and σ_X^2 are defined in terms of $P_i(\theta)$ and $Q_i(\theta)$ as before. When guessing goes high, the ITC decreases. σ_X^2 can be expressed as $[\sum \sigma_i \text{ITC}_i]^2$ (Gulliksen, 1950). Thus, high guessing would lead to decrease in coefficient α . When using the variance decomposition formula for $\text{var}(U_i) = \sigma_i^2$, coefficient α has the following relation with IRT.rel.X:

$$C_\alpha = \frac{N}{N-1} \left(\text{IRT.rel.X} - \frac{\sum \text{var}[E(U_i|\theta)]}{\sigma_X^2} \right). \quad (9)$$

When N is large, so that $N/(N-1)$ becomes nearly 1, IRT.rel.X would be greater than or equal to coefficient α in general.²

Study Design

All formulas with integrals were evaluated by a numerical integration approximation with θ ranging from -4 to 4 by $.01$ increase. The guessing level for each item was manipulated by varying g_i , in Equation 1. Other variables considered to make conditions more realistic included item difficulty (b_i), item discrimination (a_i), and TL. The TL had three levels: 20, 40, and 60. To make each item have its own discrimination, difficulty, and guessing levels, a_i , b_i , and g_i were randomly drawn from truncated normal distributions: $TN(\mu_a, \sigma_a = 0.2, \text{lower} = 0.3, \text{upper} = 1.5)$ for a_i , $TN(\mu_b, \sigma_b = 1, \text{lower} = -3, \text{upper} = 3)$ for b_i , $TN(\mu_g, \sigma_g = 1/12 = 0.08, \text{lower} = 0, \text{upper} = 0.5)$ for g_i , where TN represents a truncated normal distribution, μ is the mean of a normal distribution before truncation, and σ is the standard deviation of the normal distribution before truncation. The terms *lower* and *upper* represent the lower and the upper truncation points, respectively. μ_a was varied: 0.6, 0.9, and 1.2 for a low, medium, and high discrimination test, respectively, overall. μ_b was -1.5 , 0, and 1.5 to represent an easy, medium, and hard test, respectively, overall. μ_g was $1/12 = 0.08$, $3/12 = 0.25$, and $5/12 = 0.42$ to represent a weak, medium, and high guessing test, respectively, overall. Note that drawing from the above truncated normal distributions and varying μ makes the distributions of the sampled a_i , b_i , and g_i change approximately from positively skewed to bell-shape to negatively skewed distribution as μ increases. The total number of simulation conditions is 3 (TL = 20, 40, and 60) $\times 3$ ($\mu_a = 0.6, 0.9$, and 1.2) $\times 3$ ($\mu_b = -1.5, 0$, and 1.5) $\times 3$ ($\mu_g = 0.08, 0.25$, and 0.42) = 81. For each of the 81 conditions, 100 replications by sampling 100 times from the above truncated normal distribution were conducted. In addition to varying psychometric properties of each item, the 100 replications mimics 100 different tests in a given condition to mirror diverse tests in practice.

The coefficient α and IRT.rel.X were calculated for each replication. Thus, $81 \times 100 = 8,100$ coefficient α and 8,100 IRT.rel.X values were computed in total. To identify and summarize systematic effects of guessing in the results, the $3 \times 3 \times 3 \times 3$ full factorial analysis of variance (ANOVA) with the four factors, which are TL, μ_a , μ_b , and μ_g , was applied to obtained coefficient α and IRT.rel.X values, respectively. The purpose of this analysis is not only to assess the extent of the impact of guessing as a main effect but also to identify any interaction effects between guessing and other item characteristics. The interaction effects were investigated by examining all the interaction terms in the full factorial ANOVA approach. Given that there were a large number of observations in each of the 81 conditions (100 observations per condition), both statistical significance and effect size approaches were taken together to distinguish practically significant main effects and interaction effects. (In addition, theoretically speaking, the F test in ANOVA may not be completely reliable because of nonnormality in the distributions of coefficient α and IRT.rel.X. Their distributions in this study are negatively skewed, that is, peaked around the range between 0.7 and 1, because the conditions used here were designed to simulate actual test reliabilities found most often in practice. Note that the estimation and decomposition of sum of squares in ANOVA is not problematic regardless of normality in the data, which is the key in estimating the effect size measure exercised here. The effect size employed here was the eta square (η^2), which is the ratio of the sum of squares for an effect to the total sum of squares. Because the number of observations in the present study is balanced across all simulation conditions, the sum of squares for an effect from the full factorial ANOVA represents a unique sum of squares, and the sum of all sum of squares for all effects is equal to the total sum of squares. The eta square in this study, therefore, represents the unique amount of variation explained by the particular main or interaction effects, or the effect size controlling for all other effects considered in the study. Cohen (1988) showed the eta square values that correspond to his well-known effect size thresholds for d (effect size for a group mean difference) equal to .2, .5, and .8. His eta square classification values are .01, .06, and .14

(rounded up to two decimal points) for small, medium, and large effects, respectively, but these were based on a simple two-group comparison case only. In this study, each factor had three levels, so those eta classifications cannot be directly applied. To decide the small, medium, and large eta square values as an effect size, because the eta square is the coefficient of determination (R^2) and can be interpreted as the amount of unique variation explained by an effect in a balanced sample case, this study used .01 (1%), .09 (9%), and .25 (25%) for small, medium, and large effect sizes, respectively, which are based on Cohen's guideline for correlation classification as an effect size that includes .1, .3, and .5 for small, medium, and large effect sizes, respectively. When assessing the importance and the amount of impact of an effect, the criteria that the effect should be significant and its eta square should be at least .01 (1%) was employed to distinguish practically significant effects.

All conditions where their main and interaction effects were found to be practically significant were also expressed in terms of CTT item statistics—average item easiness (π_i) and average IRC.

Results

Before proceeding with the full factorial ANOVA for coefficient α and IRT.rel.X, the differences between them were examined. IRT.rel.X was always greater than coefficient α . The correlation between them was .99. The differences between IRT.rel.X and coefficient α ranged from .001 to .027 with the mean difference = 0.005 and the standard deviation of the differences = 0.003. Ninety-one percent of the differences was less than or equal to the difference of 0.01, and 98% of the differences was less than or equal to 0.015. Kim and Feldt (2010) also computed IRT.rel.X and coefficient α with random samples from real item response data. The differences between IRT.rel.X and coefficient α based on their report range from .002 to .006 for 30- and 25-item multiple-choice tests. Their results are within the range of the current study's simulation results here. Differences between IRT.rel.X and coefficient α values appear to be small in general.

Because the differences between coefficient α and IRT.rel.X values were small, the ANOVA results were also very similar to each other. The full factorial ANOVA result for coefficient α is shown in Table 1. (The ANOVA result for IRT.rel.X is not shown because of the similarity of the result.)

The practically significant effects which meet both statistically significant F test results under the nominal Type I error rate of .05 and eta square greater than .01 (i.e., at least small effect size or at least 1% of variation uniquely explained by the effect) for coefficient α and IRT.rel.X were TL, μ_g , μ_a , μ_b for main effects and TL: μ_g and μ_g : μ_b for interaction effects. TL showed the largest effect and item characteristic factors, μ_g , μ_a , and μ_b , showed medium effect sizes. The two practically significant interaction effects had small effect sizes and the effect size of the interaction between guessing and difficulty (μ_g : μ_b) was larger than that of the interaction between TL and guessing (TL: μ_g). Because of the overall similarity of the ANOVA results between coefficient α and IRT.rel.X, only the results for coefficient α are described henceforth. The interpretations made below for coefficient α can be applied to IRT.rel.X with little loss of generality due to the similar patterns found in this study. As noted earlier, the focus of the interpretations here is given to the guessing effect and its interaction effects.

Main Effect of Guessing

Table 2 shows the summary of the main effects. In Table 2, CTT test difficulty and discrimination are also shown by averages of CTT item easiness (π_i) and IRC that correspond to the conditions for μ_b and μ_a . For $\mu_a = 0.6, 0.9$, and 1.2 , IRC was .24, .30, and .34 on average,

Table 1. The Full Factorial ANOVA Results for Coefficient α .

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Pr(>F)</i>	η^2
TL	2	44.15	22.07	47951.07	$p < .001$	.403***
μ_g	2	22.87	11.44	24845.30	$p < .001$	.209**
μ_a	2	12.23	6.12	13288.19	$p < .001$	.112**
μ_b	2	19.78	9.89	21480.30	$p < .001$	.180**
TL: μ_g	4	1.24	0.31	674.68	$p < .001$	.011*
TL: μ_a	4	0.75	0.19	406.00	$p < .001$	.007
μ_g : μ_a	4	0.07	0.02	37.83	$p < .001$	.001
TL: μ_b	4	1.01	0.25	545.94	$p < .001$	.009
μ_g : μ_b	4	3.46	0.87	1881.22	$p < .001$	.032*
μ_a : μ_b	4	0.10	0.03	56.56	$p < .001$	.001
TL: μ_g : μ_a	8	0.03	0.00	6.80	$p < .001$	.000
TL: μ_g : μ_b	8	0.07	0.01	17.65	$p < .001$	.001
TL: μ_a : μ_b	8	0.05	0.01	14.43	$p < .001$	.000
μ_g : μ_a : μ_b	8	0.06	0.01	14.87	$p < .001$	.001
TL: μ_g : μ_a : μ_b	16	0.03	0.00	3.33	$p < .001$	.000
Residuals	8,019	3.69	0.00			

Note. The total eta square (sum of SSs [except the residual] divided by the total sum of square) was .97. SS = sum of square; MS = mean square; TL = test length. "***," "**," and "*" mean that the effect size (eta square) is large, medium, and small, respectively.

Table 2. Average Coefficient α for Main Effects.

	Small	Medium	Large
TL	20	40	60
Coefficient α	.69	.81	.86
	Weak	Medium	High
Guessing (μ_g)	0.08	0.12	0.42
Coefficient α	.85	.79	.72
	Low	Medium	High
Discrimination (μ_a)	0.6	0.9	1.2
Mean IRC	0.24	0.30	0.34
Coefficient α	.73	.80	.83
	Easy	Medium	Hard
Difficulty (μ_b)	-1.5	0	1.5
Mean π_i (item easiness)	0.83	0.63	0.42
Coefficient α	.82	.82	.72

Note. TL = test length; IRC = item-rest correlation.

respectively. For $\mu_b = -1.5, 0$, and 1.5 , item easiness, π_i , were .83, .63, and .42, on average, respectively.

It is not very surprising that all main effects considered in the study are practically significant. In addition to guessing, TL, overall item discrimination, and difficulty levels are known to influence the test reliability (Allen & Yen, 1979; De Gruijter & van der Kamp, 2008).

Figure 1. Interaction plots for TL:guessing ($TL:\mu_g$) and guessing:difficulty ($\mu_g:\mu_b$).
Note. The x = axis for difficulty represents μ_b , and the mean of π_i is in parentheses (CTT item easiness). TL = test length; CTT = classical test theory.

Increasing coefficient α was accompanied by increasing of TL and overall test discrimination. Interestingly, when the overall test difficulty was easy or medium, coefficient α on average did not change. Coefficient α decreased on average when the overall test difficulty became hard (see Figure 1 and Table 2). This particular pattern may be partly due to the interaction between guessing and difficulty ($\mu_g:\mu_b$) found here. If the amount of the impact of guessing increases as item difficulty decreases, then a hard test is affected more by guessing than an easy test. The decrease of reliability may become severe, especially as the test difficulty increases. If there is no guessing, we expect that the test of medium difficulty ($\mu_b = 0$) would show the highest reliability (given that the person ability distribution employed in this study is the standard normal distribution). In the study design, A condition where all item guessing levels are strictly zero does not exist. To investigate this case and the interaction between guessing and difficulty in detail, an additional simulation was conducted, which is explained later when the results for the interaction effects are described.

The marginal main effect of guessing on reliability, when its interactions and all effects for the item discrimination, item difficulty, and TL differences exist in the test, made coefficient α

decrease from .85 (weak guessing $\mu_g = 0.08$) to .79 (medium guessing $\mu_g = 0.25$) to .72 (high guessing $\mu_g = 0.42$). The weak guessing is a condition where a test allows minimal guessing for many items (but with a few to several items having some or high guessing). The medium guessing is a condition that one may find most often in a multiple-choice test with four options.

Interaction Effect Between Guessing and Other Variables

The practically significant interaction effects for coefficient α (and IRT.rel.X) were TL:guessing (TL: μ_g , interaction between TL and guessing) and guessing:difficulty ($\mu_g:\mu_b$, interaction between guessing and difficulty). Their effect sizes (eta squares) for coefficient α were .011 and .032, respectively, which were small. Figure 1 summarizes these two interaction effects.

The TL:guessing interaction shows that the impact of guessing increased slightly more when the TL gets shorter. In other words, the decrease of coefficient α due to the effect of guessing was larger when a test is short than when it is long. For the guessing:difficulty interaction, the impact of guessing on coefficient α increased as a test became more difficult; that is, a difficult test was affected by guessing more than an easy test regarding the decrease of coefficient α . It is interesting to see that when guessing (μ_g) was weak and medium, the medium difficulty test ($\mu_b = 0$) maintained the highest coefficient α (.87 and .83), but when the guessing was high (.42), the easy test ($\mu_b = -1.5$) achieved the highest coefficient α (.78). This explains why the main effect of difficulty (μ_b) showed the same coefficient α values for the easy and the medium difficulty test (see Table 2). To investigate the guessing:difficulty interaction with a case where all items have strictly no guessing at all, that is, $g_i = 0$, an additional simulation was conducted where TL was 40 and the difficulty (μ_b) ranged from -2 to 2 by an increase of 0.5 . One hundred tests (replications) were simulated with all other factors the same as before. Figure 2 summarizes the results from this extended guessing:difficulty interaction simulation.

It is very clear that the impact of guessing on coefficient α becomes larger as the test difficulty increases. Also, when there is no guessing in all items in a test, the coefficient α showed a symmetric pattern where the middle difficulty ($\mu_b = 0$) produced the highest value. When non-zero guessing was introduced, coefficient α was negatively affected, decreasing more as the test difficulty increased.

Difference Between IRT.rel.X and Coefficient α

When the essentially tau-equivalent assumption is violated, coefficient α is considered as a lower bound on reliability of a test in the literature (Novick & Lewis, 1967) and the amount of underestimation is not known in general. In this study, when the 3PLM is true with the presence of guessing (thereby, the essential tau-equivalence condition does not hold), the differences between IRT.rel.X and coefficient α turned out to be small in general, but as an additional interest, the differences were analyzed by the same full factorial ANOVA to investigate which factors were mostly related to the differences. (The detailed ANOVA results and plots for main and interaction effects are not presented.) All effects' p values were less than .001 except the four-way interaction term (TL: $\mu_g:\mu_a:\mu_b$, p value = .065). The effects whose eta squares were greater than 0.01 (1% of variation explained) were TL ($\eta^2 = .348$, large effect), μ_a ($\eta^2 = .302$, large effect), μ_b ($\eta^2 = .08$, small effect) for main effects, and TL: μ_a ($\eta^2 = .051$, small effect), TL: μ_b ($\eta^2 = .013$, small effect), $\mu_a:\mu_b$ ($\eta^2 = .018$, small effect) for interaction effects. For main effects, larger differences were associated with decrease of TL, increase of test discrimination (i.e., increase of μ_a), and easy or hard tests (i.e., when $|\mu_b| > 0$). The TL: μ_a interaction showed that differences between IRT.rel.X and coefficient α , due to the test discrimination differences, became larger as the TL decreased. Also, the differences due to the different test difficulty

Figure 2. Interaction plots for the extended guessing:difficulty ($\mu_g:\mu_b$) conditions.
 Note. The x = axis for difficulty represents μ_b , and the mean of π_i is in parentheses (CTT item easiness). CTT = classical test theory.

levels became larger as the TL decreased. For the discrimination:difficulty interaction ($\mu_a:\mu_b$), the impact of difficulty on the differences became larger and larger as the test discrimination increased. The difference on average was largest (.012) when the TL was short (TL = 20) and the test discrimination was high ($\mu_a = 1.2$, average IRC = .34). When this difference of .012 on average is converted into the number of items in a test using the Spearman–Brown formula, this is equivalent to about 1.3 item difference in the 20-item test, which appears to be small.

Summary and Discussion

Compared with previous research on guessing and reliability, which used rather restrictive constraints (e.g., homogeneous item difficulty, item discrimination or homogeneous inter-item correlation across items, or partial or whole parallel test assumption), the present study made a more realistic approach in studying the effect of guessing where item difficulty, item discrimination, and item guessing vary across items in a test and the parallel, essentially tau-equivalent, or congeneric assumption is not necessary. Based on the expressions of CTT item statistics, coefficient α , and IRT.rel.X, which are functions of the 3PLM parameters, a variety of test conditions were manipulated and behaviors of coefficient α and IRT.rel.X under the influence of

guessing were investigated at the population level by directly evaluating those expressions, which obviates item response data generation.

In this study, guessing in a test was manipulated by the 3PLM lower asymptote parameter, g_i . There are a few advantages of using the 3PLM approach for modeling guessing behavior in the simulation. First, each item has its own guessing level. Second, the 3PLM can be viewed as an approach to model both random guessing (where all items have the same guessing level across all different examinee levels) and partial knowledge guessing behavior (see Waller, 1989). One key insight for this matter is observed when comparing the difference in probability of a correct response between the 3PLM and the two-parameter logistic model (2PLM, which has the same discrimination and difficulty parameters as the 3PLM parameters except g_i). The difference in probability of a correct answer between the two models decreases as ability increases. The simple random guessing takes place only when the examinee's ability is low (more precisely, when the ability is far below the difficulty level). The differences in the probability of correct answer between 3PLM and 2PLM as ability goes up are mixture of random guessing and partial knowledge guessing (i.e., decreasing simple random guessing and making adjustment for partial knowledge guessing). Waller (1989) provided this interpretation of the 3PLM with supporting empirical analysis results by comparing it with other models. At least, it can be argued that the 3PLM provides an adjustment for the impact of simple random guessing and models differential impact of guessing across different ability levels. In addition to the IRF differences between 3PLM and 2PLM discussed earlier, this aspect of the 3PLM's modeling of guessing can be straightforwardly verified when evaluating the differences of item information between 3PLM and 2PLM as well. Thus, employing the 3PLM and its lower asymptote parameter with regard to modeling guessing behavior provides a simulation scenario closer to reality than does previous research, which used more restrictive constraints in studying guessing and reliability.

When a test is short, overall test discrimination is high, and overall test difficulty was either easy or hard, the differences between $IRT_{rel.X}$ and coefficient α tended to increase, but they were small in general. When applying the Spearman-Brown formula, the differences on average for the TL of 20, 40, and 60 were equivalent to losing (or adding) about two items or less.

Given the simulation conditions (TL = 20, 40, and 60 when overall test difficulty, discrimination, and guessing varied), the marginal effect of guessing was to drop reliability by about .06 (.85 - .79 = .06) between weak and medium guessing, .07 (.79 - .72 = .07) between medium and high guessing, and .13 (.85 - .72 = .13) between weak and high guessing. When using the Spearman-Brown formula, to reach .85 from .79 or to reach .79 from .72, the TL should be increased by about 1.5 times as large as the original test. If one wants to make a high guessing test have the same reliability as the weak guessing test, that is, to reach .85 from .72, the TL of the high guessing test should be increased by about 2.2 times. Guessing also interacted with TL and test difficulty. The impact of guessing was increased as the TL decreased. Zimmerman and Williams (2003), whose work was based on CTT, also reported a similar finding on the larger impact of guessing on reliability for short tests.

The guessing impact in this study was not symmetrical as the test difficulty deviated from the medium difficulty (i.e., when overall test difficulty did not match the examinee group ability mean). A hard test was much more severely affected by guessing than an easy test. When there was no guessing, the highest reliability was achieved when the test difficulty matches the ability distribution, which was the middle (or medium) difficulty test in this study. As the guessing level increased in a test, the highest reliability was achieved with a test whose difficulty is below the middle difficulty, that is, an easy test (see Figure 2). This finding is consistent with Lord's (1952) point for a condition to maximize the test reliability in terms of the distribution of item

difficulties for a multiple-choice test. The interesting observation provided in this study is that the negative effect of guessing is much less detrimental when the test is easy and it becomes very serious when the test becomes hard. Even for a weak guessing test, the impact of guessing could be grave if the test is hard (see Figure 2). For instance, when a multiple-choice test's overall guessing level is medium ($\mu_g = 0.25$), the differences between the highest coefficient α (.865 at $\mu_b = -1$) and the coefficient α values for other easy tests (.858 at $\mu_b = -1.5$ and .864 at $\mu_b = -0.5$) were .011 and .001. However, when compared with other hard tests having $\mu_b = 0.5$, 1, and 1.5, the differences were .026, .059, and .11, respectively. When a multiple-choice test's overall guessing level is weak ($\mu_g = 0.08$), the differences between the highest coefficient α (.902 at $\mu_b = -0.5$) and the coefficient α values for hard tests (.874 at $\mu_b = 1$ and 0.849 at $\mu_b = 1.5$) were .028 and .053. If using the Spearman–Brown formula, these two differences (.028 and .053) indicate that the hard tests ($\mu_b = 1$ and 1.5) should be lengthened by 1.33 and 1.64 times as long as their original TL to reach .902 reliability.

This study considered several important psychometric properties of items in studying test reliability and guessing through simulations, but there are several limitations worth noting. First, this is a simulation study and cannot mirror reality completely. For example, the ability distribution was assumed to be normal. Item difficulty, discrimination, and guessing parameters were drawn from truncated normal distributions. Although these truncated normal distributions allow nonnormal sampling of those item parameters (e.g., skewed distributions), it cannot cover all possible distributions of those parameters in reality. Especially, one might argue that item difficulty is never normal even when the mean test difficulty matches the examinee ability level, because items in educational tests typically are based on a test blueprint, which specifies the number of items in each content area that the test should cover. For further studies, variations of different ability distributions³ and difficulty distributions (e.g., almost uniform distribution or very flat normal or skewed distributions) may be tried with different types of discrimination and guessing parameter distribution forms. Second, this study is based on a parametric IRT model, 3PLM. The situation simulated here posits that the 3PLM fits the population. The use of a parametric IRF form is another limitation of the present study because in reality, even the 3PLM may not describe all possible item response behaviors. Third, as this study focused on the marginal test reliability, which is most often the practitioners' interest, the potential differential impact of guessing on reliability across different ability levels was not examined. This conditional effect of guessing across different ability levels may be further investigated in future. Last, as mentioned in the introduction, it should be noted that all analyses and inferences were made at the population level. All the behaviors of reliability and the impact of guessing described in this study are expected results at the population level. Studying the impact of guessing on reliability under finite samples, especially small samples, would be another extension of the current study.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was partially supported by the First Year Assistant Professor (FYAP) Grant from the Florida State University.

Notes

1. A recent study by Kim and Feldt (2010) also employed the item response theory (IRT) framework. However, the focus of their investigation was on IRT reliability and its comparison with other classical test theory (CTT) reliabilities, not about the systematic evaluation of the degree of the main and interaction effects of guessing and other item characteristics. Also, unlike this study, all items of a test in the simulation part of their study were generated to have the same difficulty, discrimination, and guessing values, that is, the strictly parallel item condition, which was also used in other previous studies.
2. The equality of $IRT.rel.X$ and C_α is achieved when items satisfy the essential tau-equivalent assumption, which requires that all items have the same slope, difficulty, and guessing parameter values, that is, the identical item response functions (IRFs) across all items, which is not the case in this study. See also Zinbarg, Revelle, Yovel, and Li (2005) for the relations of coefficient α with other types of reliability.
3. As per a reviewer's request, a small-scale simulation was conducted to examine the variation of coefficient α when the ability distribution is not normal. Two positively skewed and two negatively skewed ability distributions were simulated with a 40-item test. (All item parameters were drawn from the distributions explained previously and fixed across different ability distributions). The skewed distributions had 0, 0.5, or -0.5 for their means; 0.7 or 1 for their standard deviations; about 1 or -1 for their skewness; and about 1.5 or -1.5 for their excessive kurtosis. Compared with the standard normal ability distribution ($C_\alpha = .936$), those skewed distributions resulted in lowering coefficient α , which ranged from .885 to .931 when there is no guessing ($g_i = 0$). When guessing was introduced, coefficient α in the skewed ability distributions became lower (range = .77-.869). (The standard normal case was .876.) In addition, two sets of item parameter estimates from real data analyses were used with the standard normal ability distribution to calculate coefficient α and $IRT.rel.X$. Coefficient α was .833 and .854, whereas $IRT.rel.X$ was .838 and .858, respectively. Using the real item parameter estimates showed the same pattern (higher values of $IRT.rel.X$ than those of coefficient α) as in the current simulation study. Details of these additional analyses are available from the author.

References

- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. New York, NY: Brooks/Cole.
- Angoff, W. H. (1989). Does guessing really help? *Journal of Educational Measurement*, 26, 323-336.
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-479). Reading, MA: Addison-Wesley.
- Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items or between tests. *Psychometrika*, 10, 1-19.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- De Gruijter, D. N. M., & van der Kamp, L. J. Th. (2008). *Statistical test theory for the behavioral sciences*. Boca Raton, FL: Chapman & Hall/CRC.
- Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement*, 21, 347-360.
- Gross, M. E., & Wright, B. A. (1985). Validity and reliability of true-false tests. *Educational and Psychological Measurement*, 45, 1-13.
- Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: Wiley.
- Kim, S., & Feldt, L. S. (2010). The estimation of the IRT reliability and its lower and upper bounds, with comparisons to CTT reliability statistics. *Asia Pacific Education Review*, 11, 179-188.
- Lord, F. M. (1944). Reliability of multiple-choice tests as a function of number of choices per item. *Journal of Educational Psychology*, 35, 175-180.
- Lord, F. M. (1952). The relationship of the reliability of multiple-choice items to the distribution of item difficulties. *Psychometrika*, 18, 181-194.

- Lord, F. M. (1977). Practical applications of item characteristics curve theory. *Journal of Educational Measurement, 14*, 117-138.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- MacCann, R. G. (2004). Reliability as a function of the number of item options derived from the "knowledge or random guessing" model. *Psychometrika, 69*, 147-157.
- Nicewander, W. A., & Thomasson, G. L. (1999). Some reliability estimates for computerized adaptive tests. *Applied Psychological Measurement, 23*, 239-247.
- Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika, 32*, 1-13.
- Plumlee, L. B. (1952). The effect of difficulty and chance success on item-test correlation and on test reliability. *Psychometrika, 17*, 69-85.
- Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18*, 229-244.
- Waller, M. I. (1989). Modeling guessing behavior: A comparison of two IRT models. *Applied Psychological Measurement, 13*, 233-243.
- Zimmerman, D. W., & Williams, R. H. (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. *Applied Psychological Measurement, 27*, 357-371.
- Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α , Revelle's β , and McDonald's ω_y : Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*, 123-133.