RESEARCH
MEMORANDUM

SURVEY OF STUDIES ON CORRECTION FOR GUESSING AND GUESSING INSTRUCTIONS

Paul M. Freeman

February 27, 1952

Survey of Studies on Correction for Guessing and Guessing Instructions

The literature on testing was examined for empirical evidence as to the effect on reliability and validity, particularly on the latter, of correcting for guessing, and of different instructions to students regarding guessing. Since much of the early research in this area was concerned with true-false tests it seemed well to include consideration of them, even though answering true-false items undoubtedly involves processes different from those involved in answering multiple-choice items.

At the outset, it may be stated that the evidence is not conclusive. While much that is significant has been written about the theoretical need to correct for guessing, and about the psychological and instructional values of such a correction, the somewhat atomistic, or at least uncoordinated, research that has been done during the last 25 years fails to provide an answer that can be generalized widely (to include all tests given now to all people).

Tables I and II summarize the most important studies examined in the survey. The only large-scale study that systematically varied directions about guessing was that of De Graff and Ruch (Table II), who, in 1925, as reported by Ruch (12), gave a different form of an objective achievement test in history to each of ten chance groups of about 250 high-school students. The design was as follows: first, all ten groups took two equivalent forms of a recall test to be used as a criterion of validity. On a subsequent day, each pupil took an evidently factual test including the "same" items in some recognition form -- true-false, two, three, five or seven response. The five groups taking the different recognition forms were further subdivided by having a random half of each take the test with instructions to guess (when in doubt) and the other half with instructions not to guess. The tests were not speeded.

The results of this study, summarized in Table II, indicate that correcting for guessing (using the conventional formula) raised test validity, in smaller amounts as the number of alternatives increased, and had an inconsistent effect on reliability, and that instructions regarding guessing had no consistent effect on test validity. Table II a, computed from the data in Table II, shows differences between certain validity and reliability coefficients, respectively, obtained in the study.

B. D. Wood (16), in 1925, gave true-false achievement tests (probably somewhat speeded) in French and Law to groups of 100 students with instructions not to guess. He found that correlations with test and grade criteria were raised by correcting for guessing, while the correction generally resulted in lower reliability. (Table I)

The following year E. P. Wood (17) also found that the validity of true-false and 5-choice tests (probably somewhat speeded) in Government, given to college students with instructions not to guess was raised by applying the correction formula. The effect on reliability was not reported (Table I).

Gritten and Johnson (8) who administered unspeeded 5-choice vocabulary tests in 1941 to four groups of 20-30 college students in an experiment to determine the effect of confidence, found a higher validity coefficient for corrected scores ($R-\frac{W}{n-1}$) than for R or R - W.

L. L. Thurstone, in 1919, devised a formula for C, the optimal weight for Wrongs in T-F tests. Thurstone, T. L. Kelley, Brinkley and Ruch separately found evidence that the best weight for Wrongs for maximum validity was less than -1.00, probably near -.8 or -.9. Dailey (1) reported a technique for estimating the optimal weights for Wrongs in advance of calculation of actual regression weights. Data from the study of De Graff and Ruch cited above, reported by Foster and Ruch, indicated that multiple R for validity, using Rights and Wrongs, or Rights, Wrongs and Omissions (in the latter case where instructions were NOT to guess) as separate variables was greater than r between $R-\frac{W}{n-1}$ scores and the criterion (Table III). E. P. Wood's data, on the other hand, show practically no increase in validity R over the r obtained using the conventional correction (Table IIIa).

The following items from Ruch's (12) summary of his chapter "Chance and Guessing in Tests" are considered to be worth quoting in full:

6. Whether correction for chance lowers or raises reliability is still debatable. Most of the studies point toward a lowering, although one of the most extensive points toward the opposite effect.
8. The evidence is almost without disagreement to the effect that correction for chance increases the validity of test scores, especially when true-false, two response and three response tests are concerned.
10. The available evidence suggests that both more valid and reliable scores are to be obtained by instructing pupils to omit items where the answering is nothing more than a sheer guess, i.e. by using "do not guess" instructions. This does not mean that pupils should not attempt items about which they have definite "hunches" or "fringes of knowledge".
11. The two investigations (E. P. Wood and De Graff and Ruch) which studied the combined question of instructions about guessing and corrections for chance suggest that the best practice is (a) instructions against wide-spread guessing, combined with (b) correction for chance.
15. There is some evidence that the general formula $S = R-\frac{W}{n-1}$ over-penalizes in multiple-response tests, but much work is yet to be done on this problem.

A number of suggestions have been made which may be relevant to future research on guessing.

a. Jarvik (9) noted the effect on guessing of the previous proportion of responses in a given position believed by the examinee to be correct -- the "gambler's fallacy".

b. Hevner (7) found that music and art tests had higher validity and reliability

when answers were weighted according to the degree of confidence indicated by students.

c. In studies of the T-F test Fritz (3) found that approximately 60% of the incorrect answers of college students on a test containing half true and half false statements on topics unknown to the students were marked "true". Cronbach (10) pointed out that "acquiescence" (the tendency to mark an item "true" rather than "false" when guessing) impairs the validity of the conventional R-W formula.

d. Gritten and Johnson (8) tested and confirmed their hypothesis that the most valid scores are those least affected by individual differences in degree of confidence.

e. Granich (5) obtained an "index of guessing" by planting a number of "phony" questions (plausible looking but meaningless) liberally throughout the exam and informing the examinees before the testing that he had done so.

A PRIORI CONSIDERATIONS

It seemed well, in view of the inconclusiveness of the experimental evidence to examine some of the writings which are a priori or "armchair" in nature. Particularly relevant is Traxler's (14) discussion, which brings out the following points:

1 - Guessing becomes less important as the number of effective choices increases.
2 - Where there is no correction, guessing raises scores in proportion to students' willingness to gamble.
3 - There are qualitative and quantitative differences between "honest" guessing (based on hunches or half-knowledge) and "dishonest" (completely random) guessing.
4 - If alternatives are all plausible and if guesses are "honest", the correction for guessing over-corrects.
5 - When all wrong responses are not equally attractive the correction formula under-corrects.
6 - Where the examinees' task is to select "best" answers rather than "right" answers, the assumption of blind guessing is seriously challenged.
7 - When instructions are NOT to guess, examinees cannot follow them uniformly because the term "guess" itself means different things to different people, and also the meaning will vary with amount of knowledge.
8 - Instructions to answer every item, guessing when in doubt, have adverse psychological and pedagogical effects and also increase the error variance.

CONCLUSIONS

Quoting Lyerly (11), it appears that "____ the assumptions employed in correction for guessing (perfect knowledge or perfect ignorance on each item and 'pure' random guessing at unknown items) are only approximately appropriate for most objective tests. 'Partial' knowledge, positional response tendencies, the ability to eliminate one or more alternatives, the lack of independence among items or among alternatives within an item (as, for example, when alternatives are ordered along a continuum), varying degrees of 'willingness to gamble' -- Those and other circumstances weaken the validity of our mathematical model, as the results of almost any item analysis will show. The effects of these factors cannot be determined from any inspection of raw test scores, and it is doubtful that any general scoring formulas can be found which will take them into account. It may be possible however, to

devise empirical scoring methods which are approximately valid for certain kinds of tests and for certain classes of individuals."

2. Although no broad generalization can be made from the experimental work, it can be concluded that:

a) Validity of moderately speeded achievement tests given to high school and college students (27 years ago!) was increased by applying the conventional correction formula, in rough inverse proportion to the number of item alternatives.

b) The effect on test validity of instructions to guess or not to guess was indeterminate.

c) Reliability of such tests, administered with instructions to guess, was raised by applying the correction formula, also in rough inverse proportion to the number of item alternatives. When instructions were NOT to guess, the correction had a smaller effect on reliability, and it was in the opposite direction.

d) The correction formula had less effect on tests with a larger number (5-7) of alternatives per question as opposed to those with fewer alternatives (T-F 2,3 choices).

3. To find the optimal correction formula for a given test it would be necessary to determine regression weights for Rights, Wrongs and Omits. This possibly would involve an impracticably bulky procedure.

4. An important factor that has apparently not been reported as a variable in studies on guessing is that of the speededness of a test. Other variables whose variation may have an effect on guessing are age or grade level of examinees, kind of material tested, whether the items are "thought-provoking" or factual, criteria of validity, nature of directions, and probably many others. Also, examinees in the current student generation may be different in relevant ways from examinees 25 years ago. We know, at least, that they have had much more experience with objective tests.

TABLE I

Summaries of Several Studies on the Effect of Correction for Guessing on Test Reliability and Validity*

Investigators	Year Reported	N	Test	Guessing instructions	No. of choices	Reliability Unc.	Reliability Corr.	Validity Unc.	Validity Corr.	Criterion
Yerkes	1917	70	Army Alpha	?	T-F	-	-	.66	.72	Total score
					4			.84	.76	
Ruch	1924	43	Terman Group	?	T-F	.68	.56	-	-	
					5	.35	.41	-	-	
Brinkley	1924	?	History	?	T-F	-	-	.78	.82	?
					5	-	-	.76	.76	
Ruch and Stoddard	1925	Groups of 135	Achievement	?	T-F	.56	.41	-	-	
					5	.80	.77	-	-	
Paterson and Langlie	1925	111	Psychology	Do not guess	T-F	.63	.54	-	-	
Wood, B. D.	1926	100	French	Do not guess	T-F	.83	.80	.71	.75	Placement test
		100	Law			.75**	.76**	.82**	.87**	Course grades
DeGraff and Ruch	1926			(See Tables II and IIa)						
Wood, E. P.	1927	147	Government	Do not guess	T-F	-	-	.75	.84	Course grades
					5	-	-	.85	.86	
Gritten and Johnson	1941	90 approx.	Vocabulary (Form A)	Do not guess	5	-	-	.73	.81	Number right on Form B
Army Air Forces	1947	838	Numerical Op'ns	?	T-F[1]	-	-	.36	.38	Success in navigation training
					5[2]	-	-	.47	.49	
		315	Map distance	?	T-F[3]	-	-	.31	.04	Speed of Identification
						-	-	.07	.25	Mechan. Principles

*No attempt to study the effect of speed is reported in these studies. Most of the subjects were college students; Yerkes' were recruits as were those of the AAF study and Brinkley's were high-school students.

**Median of three coefficients based on results of exams in three different legal subjects

1 - Correction formula: R = 2.94W

2 - " " R = 1.98W

3 - " " R = 3W + 40

TABLE II

Condensation of Results of Ruch and De Graff's Study
1925

No. of Alternatives	Validity* (Av. 2 forms) Uncorrected	Validity* (Av. 2 forms) Corrected	Reliability (Av. 2 forms) Uncorrected	Reliability (Av. 2 forms) Corrected
7-response				
(g)[1]	.84	.87	.80	.84
(n)[2]	.90	.91	.89	.91
6-response				
(g)	.88	.91	.86	.90
(n)	.86	.89	.86	.88
3-response				
(g)	.82	.86	.84	.86
(n)	.85	.91	.89	.89
2-response				
(g)	.80	.84	.74	.86
(n)	.75	.82	.86	.84
True-false				
(g)	.74	.82	.64	.78
(n)	.76	.86	.88	.84

* Correlation with recall test

1 - (g) indicates tests taken under instructions to guess.
2 - (n) indicates tests taken under instructions NOT to guess.

TABLE IIa

Differences* in Validity and Reliability Coefficients Reported by Ruch and De Graff 1925

No. of Alternatives	Corr.-Unc. Guess (g)	Corr.-Unc. No guess (n)	n-g Unc.	n-g Corr.	g,corr.-n,unc.	n,corr.-g,unc.
			VALIDITY			
7-response	.03	.01	.06	.04	-.03	<u>.07</u>
5 "	.03	.03	.02	-.02	.05	.01
3 "	.04	<u>.06</u>	.03	.05	.01	<u>.09</u>
2 "	.04	.07	-.05	-.02	.09	.02
T-F	.08	<u>.10</u>	.02	.04	.06	<u>.12</u>
			RELIABILITY			
7-response	.04	.02	.09	.07	-.05	.11
5 "	.04	.02	.00	-.02	.04	.02
3 "	.02	.00	.05	.03	-.03	.05
2 "	.12	-.02	.14	-.02	.00	.10
T-F	.14	-.04	.24	.06	-.10	.20

* Significant (.01) differences in validity are underlined. Significance of differences in reliability coefficients was not reported.

TABLE III

Improvement of $R_{1 \cdot 23*}$ and $R_{1 \cdot 234*}$ over r_{12**} in the results of Foster and Ruch 1927

	T-F	2-response	3-response	5-response
		"GUESS"		
$R_{1 \cdot 23}$	.88	.82	.90	.93
r_{12}	.82	.84	.86	.91
Gain	.06	.08	.04	.02
		"DO NOT GUESS"		
$R_{1 \cdot 234}$	.90	.85	.94	.94
r_{12}	.86	.82	.91	.89
Gain	.04	.03	.03	.05

* 1 - criterion
2 - rights
3 - wrongs
4 - omissions

** Figures for r_{12} are taken from De Graff and Ruch's table (reported by Ruch) showing validities, corrected by the conventional formula.

TABLE IIIa

Improvement of $R_{1.234*}$ over r_{12}

(from E. P. Woods's data reported by Rush)

1927

	True-false	5-response	Completion
$R_{1.234}$	.847	.861	.890
r_{12}	.845 (R-W)	.860 (R-$\frac{W}{4}$)	.880 (R)
Gain	.002	.001	.010

* 1 = criterion
2 = rights
3 = wrongs
4 = omissions

BIBLIOGRAPHY

1. Bailey, John T. "Techniques for Estimating the Optimal Weights of the 'Wrongs' in Scoring Printed Tests" AMERICAN PSYCHOLOGIST, 1947, 2, 510 - 511 (Abstract).

2. Foster, R. R. and Ruch, G. M. "On Corrections for Chance in Multiple-response Tests", J. EDUCATIONAL PSYCHOLOGY, 18, 1927, 48-51.

3. Frits, F. M. "Guessing in a True-false Test", J. EDUCATIONAL PSYCHOLOGY 1927, 18, 558-561.

4. Gehsen, Howard W. and Kavruck, Samuel "Selected References on Test Construction, Mental Test Theory, and Statistics", Washington, U. S. Civil Service Commission, 1950.

5. Granich, Louis "A Technique for Experimentation on Guessing in Objective Tests" J. EDUCATIONAL PSYCHOLOGY 1931, 22, 145-156.

6. Greene, H. A. "A New Correction for Chance in Examinations of Alternate-response Type" J. EDUC. RES. 1928, 17, 102-107.

7. Hevner, K. "A Method of Correcting for Guessing in True-false Tests and Empirical Evidence in Support of it", J. SOC. PSYCHOL., 1932, 3, 359-362 (Abstract in Psychological Abstracts).

8. Gritten, F. and Johnson, D. M. "Individual Differences in Judging Multiple-choice Questions" J. ED. PSYCH., 1941, 32, 423-430.

9. Jarvik, M. E. "Probability Discrimination and the Gambler's Fallacy in Guessing" AMERICAN PSYCHOLOGIST, 1946, 1, 453-454 (Abstract).

10. Journal of Educational Research, XIV, 1, Feb. 1944.

11. Lyerly, Samuel B. "A Note on Correcting for Chance Success in Objective Tests" PSYCHOMETRIKA, 1951, 16, 21-30.

12. Ruch, G. M. "The Objective or New-Type Examination", Scott, Foresman and Co, 1928.

13. Ruch, G. M. et al. "Objective Examination Methods in Social Studies", Scott, Foresman and Co., 1926.

14. Traxler, Arthur E. "Administering and Scoring the Objective Test", in Lindquist, E. F., EDUCATIONAL MEASUREMENT, 329-416.

15. Votaw, D. F. and Damforth, L. "The Effect of Method of Response upon the Validity of Multiple-choice Tests", J. EDUC. PSYCHOL., 1939, 30, 624-627.

16. Wood, Ben D. "Studies ofAchievement Tests", J.EDUC. PSYCHOL., 1926, 17, 1-22.

17. Wood, E. P. "Improving the Validity of Collegiate Achievement Tests", J. EDUC. PSYCHOL. 18, 1927, 18-25.

MEMORANDUM TO: Distributees of Research Memorandums

SUBJECT: Substitution for Paragraph 3, page 4 re Research Memorandum 52-4

Paragraph 3, page 4 of Research Memorandum 52-4 should read:

"To find the optimal correction formula for a given test it is necessary to determine regression weights for Rights, Wrongs, and, in some cases, Omissions. The optimal weights for Rights and Wrongs can be approximated in advance of validation data (11a)."

The bibliography should contain the additional entry:

"11.a. Psychological Research Unit, 'Research Bulletin T44-17,' Aviation Cadet Center, San Antonio, 1944 (restricted)".

March 26, 1952
PMF:jd(jd)

Paul M. Freeman