THE JOURNAL OF EDUCATIONAL PSYCHOLOGY

THE DIAGNOSTIC VALUE OF THE WOODY ARITHMETIC SCALES: A REPLY. PART I

W. W. THEISEN AND CECILE WHITE FLEMMING

*Bureau of Educational Measurements, State Department of Public Instruction,
Madison, Wisconsin*

In the issue of *School and Society* for October 6, 1917, an article appears by Dr. Walter S. Monroe entitled "An Experimental and Analytical Study of Woody's Arithmetic Scales." In view of the extended use attained by the Woody scales during their first year and a half, this title is unusually attractive. The author of the article states that "It is the purpose of this study to learn how completely it is possible for these scales to diagnose a class, and how accurate is such diagnosis as they make." The article is characterized by an absence of any thoroughgoing attempt to show what merits the tests possess as diagnosing instruments and by an effort to prove the author's main contention that they do not diagnose through attempting to show in a rather superficial way that the Cleveland Survey tests do diagnose. The article does a grave injustice to both series of tests and to the cause of educational measurements, because it fails to reveal the real diagnostic powers of either of the tests. An untrained reader might very readily conclude that neither test possessed much diagnostic merit.

It is not the purpose of the present writers to attack the statement that the Cleveland Survey tests can be used to diagnose. In fact the great number of examples of various types which they contain make them quite comprehensive and useful for the purpose whenever sufficient time is available for examining fully the symptoms revealed. We shall however in the first part of this paper attempt to answer the principal criticisms which Dr. Monroe has made of the Woody tests. In the second we shall endeavor to point out some of the possibilities of the tests as diagnosing instruments.

I

In the first place, to diagnose means more than to find class scores and to compare them with established standards. The article referred to above treats diagnosis in this limited sense. So long as a diagnosis is only in terms of standards of attainment, it is superficial. It must go further. It must reveal the particular processes in which pupils fail, not necessarily the examples. It must reveal the particular mental processes or habits at work which are the seat of the difficulty. When a third grade child adds example 3 of Woody's addition test 17 and obtains 10 as the answer, it is of

2

little service to treat this merely as an error. The fact to be considered is that the child has added two, seven, and one, and obtained ten. This habit must be broken down and the habit of two-column addition built up before the child can be expected to progress satisfactorily. In any case a diagnosis must seek to reveal the mental processes at work tending to produce failure. Hence we must not be led astray by assuming that Dr. Monroe's exhibit of scores made on the Woody and the Cleveland Survey Tests represent the diagnosing merits of either test. Some conception of the diagnosing merits of the Cleveland Survey Tests may be had by reference to Counts' study.*

Monroe contends quite correctly that "arithmetical ability" consists not of a single ability, but several abilities. He points out for instance that Ballou records 14 types of examples in the addition of two fractions. This makes way for his first criticism of Woody's tests, namely, that they do not include all important types of examples required to test these several abilities. In reply it may be said that while the greater number of types may be admitted, it does not necessarily follow that children must be tested on all of them before fundamental weaknesses in the work of a class can be discovered. The tests do not contain all types and it is not essential to a working diagnosis that this be the case. A working diagnosis must furnish clues as to the most probable causes of failure—clues which can be used as starting points for remedial measures. This the Woody tests do. When a teacher discovers that a number of pupils in attempting to add fractions where the

*GEORGE S. COUNTS. *Arithmetic Tests and Studies in the Psychology of Arithmetic*, University of Chicago Press, Supplementary Educational Monographs No. 4, 1917.

addition is indicated by the + sign, add numerators to numerators, and denominators to denominators, or that some add numerators to denominators, it is not necessary to extend the testing to all types of examples in fractions. In one small class of nine sixth grade pupils tested by Woody's Addition Scale, Series A, six added numerators to numerators and denominators to denominators, in each of the six examples in addition of fractions where the addition is indicated by the plus sign, *e.g.*, in example 27, $\frac{1}{8} + \frac{1}{4} + \frac{1}{2} =$, the answer $\frac{3}{14}$ was obtained through such a process. In this school the course of study calls for addition of fractions, including the types indicated by these examples, in grade five. The results of the test are good evidence that the subject was not well taught, and the prevailing type of error indicates the nature of the remedy.

Furthermore, we must not lose sight of the fact that a satisfactory test must have commendable administrative features as well as supervisory and psychological values. While a test must afford a working diagnosis of teaching needs, at the same time it must be such that it can be readily administered. For that reason it cannot contain an unduly large number of examples. The inference which the writer leaves us to draw, namely, that the Cleveland Survey tests provide more types of examples than do the Woody, is not founded upon fact, as one may see by a careful comparison of the two series of tests. All things considered, Woody has combined the essentials of a good test and scale to a remarkable degree. To meet all the requirements that Mr. Monroe is asking would necessitate tests of unwarranted length and tests which would require more labor to summarize properly and interpret than most school systems can afford. The four Woody tests require eighty minutes as it is.

The criticism of the tests on the ground that the diagnosing power of a test is limited to the types it contains is rather unjust. By following up the test with slightly different examples of about the same degree of difficulty, one may discover other weaknesses than that expected from a given type of example itself. Furthermore, in all testing a basic assumption is that the examples or questions selected represent a random selection of the entire field. If correlation studies have shown us anything, they afford grounds for believing that if pupils are weak in particular types of examples, they are very apt to be weak in types somewhat similar. Hence

when wrong or confused mental habits are discovered, it is not necessary to test on all types before we can predict certain types on which they would probably fail.

Dr. Monroe makes the statement that "a satisfactory diagnosis must either be complete, or, if not, its limits must be known. Otherwise the mistake will be made of assuming that a partial diagnosis is complete." In reply we may say first, that we know something of the limitations of the test when we know that it does not contain all types or ramifications of types of examples. Nowhere does Woody claim that his tests will furnish a complete diagnosis. He does say that they will "aid greatly in locating the weaknesses of the class,"* a statement not out of keeping with the truth. Second, a satisfactory diagnosis need not necessarily be complete in all details. A diagnosis is satisfactory which reveals a number of specific class weaknesses or incorrect habits formed which are causing errors. The teacher desires to know where to attack in order to bring about improvements. She does not attempt to remedy all weaknesses at once. As we have said the most serious of them furnish her with starting points. If she be at all competent she will make use of additional examples somewhat similar in nature, and difficulty in order to discover other limitations of the class.

For the purposes of diagnosis it is not so important to know how many errors were made on particular examples as it is to know what is causing the errors. There is a vast difference between errors as such and recurring errors of a similar nature, caused by the same mental processes or habit. This is a distinction which the writer of the article apparently disregards when he discusses the diagnostic features of tests. When several pupils in a class give the same incorrect answer to an example, or when several errors of a similar nature are made by an individual, the difficulty can usually be traced to a single source, *e.g.*, in answer to example 16 of the subtraction test 270 it was found that 120 was given by

190

a number of children in several classes in one city. How to account for it was for a time a puzzle. It finally developed that it was probably due to an insufficient familiarity with the Austrian method of subtraction. The mental processes involved after putting down the cipher was somewhat as follows: "Seven and what number

*CLIFFORD WOODY. *Measurements of Some Achievements in Arithmetic*, p. 23.

make nine? Put down two. One and what number make two? Put down one." Example 18 of the subtraction scale 1000 was used
537

usually solved similarly by the same pupils. It was discovered that these children in the second and third grades had been taught to subtract by the common method of borrowing. The present teachers of these children, now in grades four and five, use the Austrian method, teaching subtraction as a form of addition. Hence the confusion. Now to have treated the answer 120 to example 16 merely as an error would not be diagnosing an evident need of this class. To discover the mental process back of the error is to diagnose. The number of such mental processes causing errors which can be discovered through the use of the Woody tests is sufficient to furnish reasonably satisfactory diagnosis of teaching needs.

A third objection is raised on the ground, that Woody, as Monroe claims, has used the statistical rather than the analytical method of selecting his examples. As a matter of fact, Woody has used a combination of the two methods. We wrote to Dr. Woody for a definite statement on this point. In a reply he says "The problems for my scales were chosen primarily for the purpose of analysis of errors made by pupils."* Statistical considerations were secondary. Nearly every example was selected because it added some new feature not contained in others already chosen. The examples rejected because of their failure to meet certain statistical requirements were few as any one may discover by correspondence with Dr. Woody himself. In this letter he states that "Those problems which were rejected had a particular characteristic which made them more or less local in their nature. By local I mean such a peculiarity that a class in the third grade would understand the meaning of such a problem while a class in the fourth would not." He says further with reference to the statistical selection of examples in Series "A"—"The steps were not equal between the different problems. This is due to the fact that the problems were arranged in order of difficulty and given that particular order because I did not wish to throw out problems merely from a statistical point of view. I preferred to retain the problems because of their diagnostic value rather than throw them out because they did not make uniform steps which a statistical scale would be inclined to do."

*Letter of January 23, 1918.

One decided advantage gained when we select examples in order of increasing difficulty, as we do when we use the statistical method, is that we obtain a graded series. The advantage of such a series in determining achievement in arithmetical abilities is one which should not be overlooked. This feature of Woody's scales resulting from the use of the statistical method is one for which no other set of tests in fundamental operations provides to a like degree, and it is one whose value Dr. Monroe fails to consider at all in his criticism of the statistical method of selecting examples.

Another objection to the statistical method of selecting examples is raised by the writer of the article because only those examples were chosen which were solved by a gradually increasing percentage of pupils from the lower to the higher grades. The writer states that "Under normal conditions the per cent. of pupils doing the examples correctly should not increase gradually from the lower to the higher grades." He does not take into account the fact that a gradual increase in the sense used by Woody means that *beginning with the grades in which some of the pupils are able to solve a given example there should be a gradual increase in the percentage of pupils in each grade above who solve it correctly and not a decrease from one grade to the next.*

Moreover, Woody attached no weight to scores made on examples solved by less than 2.15 per cent. of the children in the grade (*i.e.*, at a distance greater than 3 P. E. from the median of the distribution for the grade) in locating the examples on the scales. Examples solved by more than 2.15 per cent. but less than 25 per cent. (*i.e.*, between 1 P. E. and 3 P. E. from the median) were given only half weight. Thus no one need be concerned about the importance attached to examples much too difficult for children of the grade in making up the scales.

The writer of the article states that "Before a type of example is taught a pupil can not be expected to possess the ability to do it." Nevertheless, much to our surprise, the child sometimes does. But however that may be, we know of no one who contends that he should. Certainly Woody does not, for nowhere in his monograph do we find a statement to the effect that he does. Monroe says further that "at the time in his school career when it is taught he will learn more rapidly than afterwards." He refers to the median achievement of pupils on the Cleveland Survey Tests as tending

to support his statements. For the benefit of the reader the table showing these scores is reproduced here, Table I. Dr. Monroe calls particular attention to the scores on examples in fractions.

Tests H and O as used in the Cleveland Survey are tests in fractions. These figures indicate no score for grades 3 and 4 on test H and for grades 3, 4, and 5 on test O. Evidently the type of example included in test H was first taught in grade 5 and that of test O in grade 6. Obviously the changes in score from 0 in grade four to 5 in grade five on Test H, and from 0 in grade five to 3.1 in grade six on Test O, appear large. But this change can not well be in-

TABLE I
*Medians in Each Arithmetic Test for All Grades—Cleveland Survey**

Test	Grade					
	3	4	5	6	7	8
A	13 4	17 8	22 2	24 8	26 7	27 5
B	9 3	13 4	17 2	19 8	21 5	26 0
C	6 5	12 0	15 5	16 6	17 7	19 0
D	6 3	12 4	15 7	18 5	20.8	22 5
E	4 3	5 3	6 3	6 8	7 5	7 8
F	2.0	4.9	6 7	7 5	8 6	10 1
G	2 0	3 9	5 2	5 5	5.9	6 6
H	0 0	0 0	5 0	5 5	7 7	8 5
I	0 6	1 1	2 0	3 1	4 0	4 7
J	1 9	3.2	4 0	4 4	4.9	5 7
K	0 0	4 0	6 8	8.5	10 1	12 5
L	0 0	1 7	2 5	2 8	3 2	3.9
M	1 4	2 5	3 2	3 8	4 4	5 1
N	0 0	0 8	1 3	1 7	2 0	2 6
O	0 0	0 0	0 0	3 1	4 1	5 5

terpreted as a "gain from grades four to five" much larger than the improvement between subsequent grades. The figures indicate rather a change from a condition of *no ability* before the examples have been taught to a condition of *ability* resulting from the teaching. After this ability appears there is an uninterrupted increase throughout the succeeding grades for both Test H and Test O. Moreover Table I shows, for nearly all of the other tests, a similar regularity of improvement from grade to grade. This may be seen graphically for a number of the tests from Figures I and II. The curves for tests E, F and I, shown in Figure I, represent tests in addition,

*From CHARLES H. JUDD. *Measuring the Work of the Schools*, Table X, page 98.

subtraction and short division respectively. Test E is composed of single column addition examples of five numbers. Test F includes examples of the type 1269, and Test I of the type

772

6)57432. In none of these is there anything but a gradual improvement from grade to grade. Moreover, were one at first thought inclined to accept Monroe's assertion that "at the time in his school career when it is taught (*i.e.*, a given type of example) he will learn more rapidly than afterwards," he would have evidence from Test F that the statement is far from true. The median score for the fourth grade is more than double that of grade three.

In Figure II, tests L, N and O are represented. Test L is a test in multiplication of four place numbers by two place numbers. Test N is a test in division of the type 48)28464, and test O is a test in fractions. The curves for these tests confirm the evidence of gradual improvement shown in figure I. Certainly the facts of Table I from the Cleveland Survey Tests cited by Dr. Monroe do not offer satisfactory evidence to support his objection to Dr. Woody's statistical method of selecting examples. Still another objection is offered to the Woody tests on the ground that too few of the fundamental combinations are included. He asserts that the addition test, for example, includes only two examples in fundamental combinations. The addition scale has in effect not only two examples, but six that offer pupils an opportunity to exhibit their knowledge of the fundamental combinations. All but number two of the first seven examples, call for fundamental combinations. As may be seen they are additions of two numbers without carrying.

(1)	(3)	(4)	(5)	(6)	(7)
2	17	53	72	60	3+1 =
3	2	45	26	37	

Though an example may be of two columns, it serves to diagnose ability in the fundamental combinations so long as carrying is not required, *e.g.*, example 5, 72 gives a child an opportunity to exhibit

26

his knowledge of the combination 6 and 2, and that of 2 and 7. In the subtraction test there are ten examples, numbers 1, 2, 3, 4, 5, 8, 9, 10, 11, 12, that are essentially fundamental combinations. These are:

8	6	2	9	4	59	78	7-4 =	76	27
5	0	1	3	4	12	37		60	3

FIGURE I

FIGURE II

In the multiplication test the first eight examples are serviceable for measuring ability in multiplication without carrying.

$$\begin{array}{ccccccc}
 3 \times 7 = & 5 \times 1 = & 2 \times 3 = & 4 \times 8 = & 23 & 310 & 7 \times 9 = 50 \\
 & & & & 3 & 4 & 3 \\
 & & & & \text{---} & \text{---} & \text{---}
 \end{array}$$

Eleven of the first twelve division examples are of the simplest kind. They include division by 1, 2, 3, 4, 6 and 9 without remainders or carrying.

$$\begin{array}{ccccccc}
 3 \overline{)6} & 9 \overline{)27} & 4 \overline{)28} & 1 \overline{)5} & 9 \overline{)36} & 3 \overline{)39} & 4 \div 2 = 9 \overline{)0} \quad 1 \overline{)1} \\
 6 \times \dots = 30 & & 2 \div 2 = & & & &
 \end{array}$$

It will thus be seen that the tests offer many more opportunities for diagnosing abilities in fundamental combinations than appear at first glance. The truth is that the tests and the performances upon them require careful study to discover all that they can reveal. Not many examples could be added to the tests in their present form without necessitating the elimination of others. Monroe in demanding numerous additions to the tests strangely enough fails to indicate any desirable eliminations. To increase materially the number of examples on the test would as we have pointed out elsewhere make the use of the tests a very laborious and burdensome process.

The writer of the article bases his objection to the test of ability in fundamental combinations made by the Woody addition scale on the ground that pupils who made a perfect score on the example 2 made less than average scores on the Cleveland Survey Test A, 3 consisting of examples in fundamental combinations. But — what did these same children do on examples 3, 4, 5, 6, and 7 of the Woody addition test, we ask? Were they able to make all of the fundamental combinations in these as well? Monroe states “that the Woody Scale revealed no differences in the ability of these pupils in using the addition combinations. The Cleveland Survey Test A reveals quite striking differences in abilities.” But is it true that there were no differences? Where are the facts on examples 3, 4, 5, 6 and 7?

Dr. Monroe raises the further objection that the addition test

contains only one example of the type of example 18 which is 2563
 He reports a ninety-three per cent. score on this example by a 1387
 seventh grade class which when tested by the Cleveland Sur- 4954
 vey Test M scores below average. Woody has at least six oth- 2065
 er examples of two or more columns in more than two num-
 bers, requiring carrying, as follows: ———

(11)	(13)	(15)	(16)	(17)	(22)
32	23	100	9	199	547
59	25	33	24	194	197
17	16	45	12	295	685
—	—	201	15	156	678
		46	19	—	456
		—	—		393
					525
					240
					152
					—

The critic fails to record what this class did on these examples. The same type of ability required in the first three columns of example 18 is represented in example 17. These two examples together with example 15 would give a more accurate idea of the ability of the class to add and carry than 18 alone. To the examples above might be added examples 19, 20, 21, 24, 31 and 33, which, though they involve decimals and United States money, may also be used to diagnose ability in addition. These examples are:

(19)	(20)	(21)	(24)	(31)	(33)
\$.75	\$12 50	\$8 00	4 0125	113 46	49
1.25	16 75	5 75	1.5907	49 6097	28
.49	15.75	2.33	4.10	19.9	63
—	—	4 16	8.673	9 87	.95
		94	—	.0086	1 69
		6 32		18 253	.22
		—		6.04	.33
				—	36
					1 01
					.56
					.88
					75
					56
					1.10
					.18
					.56
					—

Apparently the writer of the article has singled out a few examples and recorded not causes of error but rights and wrongs on these few examples, forgetting that there are others in the test. He would have us use the Cleveland Survey Tests which contain but a very few types of examples in addition. They contain one test in simple combinations, two in single column addition, one in addition of similar fractions, one in addition of examples similar to Woody's 18th, and one containing examples in addition of dissimilar fractions. They include no examples in decimals, U. S. money, denominate numbers, or mixed numbers. Without any thorough-going attempt to reveal the diagnosing powers of either test he assumes to relegate one test to the scrapheap and to elevate the other upon a pedestal.

Finally, he points to the subtraction combinations on the Woody Scale, nine in number, and to examples 16, 17 and 18, which are:

270	393	1000
190	178	537
<hr/>	<hr/>	<hr/>

as being insufficient to reveal weaknesses in these phases of the arithmetical work. In fact he tries to show that the results attained with them may be materially different from those obtained on similar but longer tests. To substantiate this position he resorts to a comparison of performances made by a single class from each of grades 4 to 8 inclusive on these examples and on the Cleveland Survey Tests B and F which consist of similar examples. In the first place the number of pupils tested was too few to warrant any safe conclusions. Second, he records the average *per cent.* of examples correct on those from the Woody Scale and the average *number* of examples correct on the Cleveland tests. The latter he compares with the standard average on the Cleveland tests, but he fails to record the standard *per cent.* of examples correct for these examples of the Woody Scale. Such computations obtained by entirely different methods are not comparable. In order to make any sort of comparisons between the two tests from this standpoint, the standard *per cent.* of examples correct should be calculated. Then it will be possible to say that a given class was below standard as judged by one test or above as judged by the other. From Dr. Monroe's statements one is led to believe that children who scored high on the Woody examples did not do well on the Cleveland tests. For he says of the subtraction combinations, "According to

the Woody Scale the pupils do these examples very satisfactorily * * * * according to the Cleveland Survey Test these pupils are below standard excepting in the sixth grade, and there they are just up to standard." Again he says of examples 16, 17 and 18: "It should be noted that according to Test F the seventh and eighth grade classes are below standard in ability to do this type of example. The Woody scale shows these pupils to stand very high in ability to do examples of this type."

Now are the impressions intended here in keeping with the facts? Table II gives the average and standard number of examples correct on the Cleveland Test B and the average and standard per cent. correct on the examples from the Woody Scale. The last of these are computed from the tables in Woody's monograph.* The others are taken from the tables given by Monroe.

TABLE II

Showing the Average and Standard Number of Examples Correct on the Cleveland Survey Test B and the Average and Standard Percent of Examples Correct for Examples 1, 2, 3, 4, 5, 6, 7, 10 and 13 of the Woody Subtraction Scale

	IV	V	VI	VII	VIII
Average Cleveland Survey Test B. . .	10 6	12.7	19 3	18.3	18.
Standard average survey test B.	11.3	16.	18 3	21 4	24 4
Average per cent of examples correct					
Woody Scale.	96.	99.	99.	97.	96.
Standard per cent of examples correct					
Woody Scale.	92.9	97.4	98 6	98.8	99.4

TABLE III

Showing Facts Similar to Table II for the Cleveland Survey Test F and for Examples 16, 17, 18 of the Woody Subtraction Scale

	IV	V	VI	VII	VIII
Average Cleveland Survey Test F . . .	1.6	4 9	7.4	7.7	7 2
Standard average survey test F.	3 9	6 4	7.3	9.	10.2
Average per cent correct Woody Scale	54.	76.	92.	86.	90
Standard per cent correct Woody Scale	49.	79.	87.	93.	95.

*Ibid., Tables XI and XXII. The order and numbering of examples in Table XI is that used by Woody in collecting the original data and differs from the present order on the scale. Table XXII furnishes a key to the corresponding orders.

It is to be observed from Table II that when measured by either test grade six is above standard and grades seven and eight are below. Thus for these grades the two tests do not point to different conclusions. According to Table III,* grade six is again above standard on both tests and grades five, seven and eight are below standard on both tests. This is far different from what the writer would have us believe. Even from his own figures, derived as they are from but one class in each grade, the tests are shown to point in the same direction. The assertion made that "The Woody Scale shows these pupils (*i.e.*, grades seven and eight) to stand very high in ability to do examples of this type" does not accord with the facts. The two tests do diverge in grades four and five of Table II and in grade four of Table III. Two possible explanations may be offered. Tests given to more than 7000 Wisconsin children indicate that Woody's standards for each of his tests are too low in grades four and five. Standards in both tests probably need refining. Then the element of speed may be more effective in these grades. The averages on the Cleveland tests are based upon scores made in thirty seconds. They do not represent the ability to do examples of a given type, but the ability to do a given number of them in thirty seconds, two quite different things. Furthermore, one might well question the validity of results obtained from a limited number of pupils in so short a time. An error of ten per cent. in either direction in the time actually taken by a class is not impossible or unusual. From the facts which Monroe has given us we must conclude that the writer of the article has failed to substantiate his claim that the Woody test in subtraction contains an insufficient number of examples of the type just discussed, and that conclusions drawn from it may be misleading.

*The standard per cent. correct for the Woody examples is computed from Table XXIV of Woody's monograph. See his Table III column headed "Subtraction," and Table XXV column "Final Value," for the key to corresponding orders.

[(Part II will appear in the December number.)]