DOCUMENT RESUME

ED 312 291 TM 014 053

AUTHOR Albanese, Mark A.; Jacobs, Richard M.
TITLE Reliability and Validity of a Procedure To Measure Diagnostic Reasoning and Problem-Solving Skills Taught in Predoctoral Orthodontic Education.
PUB DATE 88
NOTE 13p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Clinical Diagnosis; Cognitive Processes; *Construct Validity; *Dental Students; Dentistry; Higher Education; Medical Education; *Medical Students; Multiple Choice Tests; *Problem Solving; Professional Education; Test Reliability; *Test Validity
IDENTIFIERS *Diagnostic Skills; Internal Consistency; *Reasoning Tests

ABSTRACT

Preliminary psychometric data assessing the reliability and validity of a method used to measure the diagnostic reasoning and problem-solving skills of predoctoral students in orthodontia are described. The measurement approach consisted of sets of patient demographic data and dental photos and x-rays, accompanied by a set of 33 multiple-choice items with from 2 to 10 options. Students were only able to complete two exercises in a 50-minute testing period, so that content specificity of any sizable magnitude might be a problem. Members of a second-year dental school class were divided into two groups, of 35 and 33 students, to take two versions (each containing two different cases) of the examination for a total of four different cases. Prescriptive items showed uniformly higher internal consistency reliability estimates than did the descriptive items. The intercorrelations among data from the same test administration, correlations between scores for different problems at different testing times, and correlations of class attendance with test scores were all sufficiently large to support the construct validity of the measurement procedure. The metnod appears to offer a viable method of assessing problem-solving skills in orthodontics, with potential for uses beyond undergraduate education. Three tables present reliability data. (SLD)

RELIABILITY AND VALIDITY OF A PROCEDURE TO MEASURE DIAGNOSTIC REASONING AND PROBLEM-SOLVING SKILLS TAUGHT IN PREDOCTORAL ORTHODONTIC EDUCATION

Mark A. Albanese, Ph.D.
Richard M. Jacobs, D.D.S., Ph.D.
The University of Iowa

RELIABILITY AND VALIDITY OF A PROCEDURE TO MEASURE DIAGNOSTIC REASONING AND PROBLEM-SOLVING SKILLS TAUGHT IN PREDOCTORAL ORTHODONTIC EDUCATION

MARK A. ALBANESE, PH.D.
RICHARD M. JACOBS, D.D.S., PH.D.
THE UNIVERSITY OF IOWA

Objectives:

To assess the reliability and validity of a procedure developed to measure diagnostic reasoning and problem-solving skills taught in predoctoral orthodontic education.

Perspectives:

Measuring diagnostic reasoning and problem-solving skills of health science students has proven to be a very intractable problem. Among various approaches used by researchers to measure such skills have been patient management problems (McGuire and Babbott, 1967) and computer simulations (e.g., PLATO). These approaches have achieved limited success. A major difficulty has been arriving at a score that can be considered to represent problem-solving skills beyond those assessed in a multiple choice item test (see for example Webster et al., 1988). Another persistent problem has been content specificity i.e., the apparent dependence of student performance in any given problem-solving situation on the particular content being assessed (Elstein et al., 1978). Thus, a fairly large number of different simulations must be administered in order to obtain a representative sample of students' general problem-solving behavior.

In recent years, more progress has been made in assessing problem-solving skills. The Medical Reasoning Aptitude Test (MRAT) has shown promise as a tool for assessing student general problem solving-skills for admission considerations (Vu et al., 1987). The Objective Structured Clinical Exam (OSCE) and the new Computer Based Exam (CBX) developed by the National Board of Medical Examiners have also produced encouraging results. From a practical standpoint, however, the last two approaches are quite complex; and the MRAT was developed as a general aptitude measure. OSCE's require a significant effort in training raters, setting up equipment, and having adequate staffing available. The

new CBX materials require a sophisticated computer-driven videodisc system. It would seem useful, therefore, to develop a problem-solving achievement measure that is much less complicated and expensive to administer and score, and which can be administered in a large group setting.

The particular problem-solving situation that served as the focus of this study is predoctoral orthodontic education. Skills required to assess the complexity of a particular orthodontic problem are fairly intricate. They involve a great deal of analysis, synthesis and reconciling of multiple morphologic variables and diverse patterns of dental, facial and skeletal structures. As a rule, developing an acceptable plan of orthodontic treatment calls for hard decisions in regard to extraction of some teeth and direction of tooth movement. These decisions rely upon complex forward reasoning (Chase and Simon, 1973) which typically is not taught in undergraduate orthodontic education. It would seem that the logic of diagnosis and orthodontic problem solving has not been articulated at this level because of the prevailing belief that diagnostic reasoning and problem-solving skills can evolve only within the framework of actual clinical practice combined with tutorial mode of learning.

At our institution over the past several years, alternative instructional strategies were developed in an effort to teach diagnostic reasoning to large classes of predoctoral dental students (Jacobs, 1987). These instructional methods are based on repetitive practice in application of facts and principles to real life orthodontic problems. The teaching strategies used included encouragement of student participation, positive reinforcement, modeling, systematic enhancement of cues, and the use of feedback--corrective procedures. Because this was a fairly dramatic departure from a conventional mode of instruction, it seemed important to create a procedure for evaluating how well students problem-solving skills were developed.

Methods

This study describes preliminary psychometric data assessing the reliability and validity of the method used to assess the problem-solving skills of students exposed to the

newly developed instructional methods. The measurement approach consists of stimulus material accompanied by a standard set of 33, 2- to 10-option, multiple choice items. The stimulus material includes patient demographic data; slide projections of patients' face, teeth, gums, and surrounding structures; and tracings of x-rays of the face and mouth region. There are 12 items that require identification of the descriptive parts of the problem, depicting the scope of malocclusion, followed by 21 items of a prescriptive nature, depicting the choice of treatment approach and strategy. The 12 different descriptive characteristics and the 21 different prescriptive options remain constant across cases; only the stimulus materials change. Thus, it is a very flexible procedure that can be applied to almost any orthodontic problem. Also, since it is based upon slide projection and easily reproduced tracings, it can be given to large classes of students in a single administration.

Students are able to comfortably complete a single problem-solving exercise in approximately 25 minutes. A test administered in one 50-minute period would only be able to include two such exercises (referred to as cases hereafter). With only two cases represented on a test, content specificity of any sizable magnitude could be expected to pose a major problem.

In order to determine to what degree the measurement procedure was affected by content specificity, data were collected immediately post-course (test 1) and again one week post-course (test 2), at the conclusion of the 1988 administration of the course. Random halves of the class (group 1: N = 35 and group 2: N = 33) were assigned different forms at the test 1 administration. Each of the two forms (A and B) contain two different cases. For the test 2 administration, the alternate form of the test was administered. Thus, over the two forms, there were a total of four different cases. Group 1 received form A at the test 1 administration and form B at the test 2 administration. The reverse was true for group 2.

Internal consistency reliability of case scores, correlations between scores obtained on different cases administered at the same time as well as correlations across testing

periods and cases were computed. Since attendance was considered to be a critical part of the learning experience, as class time was spent on practicing systematic problem solving on real life orthodontic cases, record of attendance during each class period was maintained. Therefore, the number of class periods attended was correlated with test performance, as an estimate of the construct validity of the measurement procedure.

Data source:

Sixty-eight second-year dental students enrolled in a required problem-solving course in the summer of 1988.

RESULTS

Table 1 shows the means, standard deviations and internal consistency reliability estimates (alpha) for each case and test administration. The prescriptive items showed uniformly higher internal consistency reliability estimates than did the descriptive items. Differences in reliability estimates between the prescriptive and descriptive subtests for the same cases ranged from a low of .19 to a maximum of .66. This is potentially related to the greater number of items in the prescriptive test (21 versus 12). However, test length differences do not account for all of the difference because adjusting the reliability estimates of the descriptive subscore for test length differences using the Spearman-Brown Prophesy formula still yielded values substantially lower than those of the prescriptive test. (The range in differences in reliability estimates for the prescriptive and descriptive subtests was reduced to .09 to .58 after adjusting for differences in test length). The most likely reason for the lower descriptive subscore reliabilities is the proportionately smaller standard deviations. Whereas the prescriptive subscore standard deviations covered from 11% to 20% of the score range, the descriptive subscore standard deviation only covered from 4% to 11% of the score range. These proportionately smaller standard deviations are likely to be due to the ceiling effect of having an average of 96% correct on the descriptive subtest versus 83% correct on the prescriptive subtest.

Table 2 shows the raw and corrected (for attenuation) intercorrelations among the cases obtained from the same test administration. Note that the eight raw correlations range from .28 to .83, with all but one at or above .58. Combining the prescriptive and descriptive subscores together yields inter-case correlations ranging from .41 to .85. These seem to reflect relatively high correlations for a case analysis type problem, since it is common in the patient management problem literature to find correlations in the .1 - .3 range (Norman et al., 1983). When these correlations are corrected for attenuation, they increased markedly. This was especially true for the descriptive subscores in which the corrected values all exceeded 1.0.

Table 3 contains the correlations between scores obtained for different case problems obtained at different testing times. In essence, they are test-retest alternate form reliability estimates. The values are again relatively high ranging from .38 to .78 for the descriptive subscores (mean = .60) and from .37 to .79 for the prescriptive subscores (mean = .65). One perplexing finding is that although the internal consistency reliability of the descriptive subscore was substantially lower than that of the prescriptive subscore, the between case correlations were comparable and sometimes higher for the descriptive subscore. This is also reflected in the disattenuated correlations for the descriptive subscore being in excess of 1.0. As noted earlier, the discrepancy may relate to the comparatively high mean scores and small standard deviations of the descriptive subscores (see Table 1). Alpha tends to need large variances to achieve high values. It may be that for these data coefficient alpha is a less appropriate estimator of score reliability than is a test-retest correlation.

The correlations of attendance with the test scores produced values that ranged from .06 to .43 with a mean of .28 for the descriptive subscore. Similarly, the prescriptive subscore correlation ranged from .12 to .32 and had a mean of .27. A majority of the individual correlations were statistically significant at the .05 level or approached it at the

.10 level. These values were sufficiently large to be interpreted as supporting the construct validity of the measurement procedure.

DISCUSSION

The experimental testing method has a number of features that are appealing. The use of visual and case presentation information as stimulus material should provide a relatively realistic testing situation. It is also a very flexible testing procedure in that to change the case, one need only to change the stimulus material. This is somewhat similar to the item shell concept sometimes used as part of a domain-referenced testing system (Hively et al., 1973). Besides its realism and flexibility in construction, the test can be administered in large groups, making it even more appealing as an evaluation tool.

One drawback of the testing procedure, however, is that each case takes a relatively long time for students to complete (approximately 25 minutes). Thus, only a very limited number of cases can be administered at a given time. If performance on different cases is markedly different, as has been found with patient management problems, this would be a substantial weakness. The results of this study suggest that there is relatively good consistency across cases. This would further suggest that a relatively "content-free" set of problem solving skills are assessed.

The reason for this content generalizability compared to the content specificity found with other procedures used to assess problem-solving skills may be due to a number of characteristics of the experimental method. A major reason may be that it focuses on a more limited content area. Unlike patient management problems which have the potential to involve multiple organ systems, orthodontics is, by its very nature, limited to a somewhat narrower focus. With fewer content options open for consideration, the problem-solving process may be more stable across cases. A second reason for the stability probably relates to the problem-solving algorithm taught during the orthodontic course. The algorithm, utilizing a process of modeling and converting diagnostic variables, was developed to work across varying types of orthodontic problems (see Jacobs, 1987, for a description of the

algorithm employed). Another reason for the potential stability of the results is that the students were in their first year and had very little prior learning of complications and exceptions to the rule to distract them from faithfully applying the algorithm.

At the very least it can be said that the experimental procedure offers a viable method of assessing problem-solving skills in orthodontics. Presently it has only been used in undergraduate education, but it might also be very useful for licensure examinations. It offers the potential for application to areas besides orthodontics, but the extent of its transportability will need to be determined with each application. In these applications, whether the experimental procedure offers a procedure freer from content specificity problems found with patient management problems will await the results of future research.

TABLE 1

MEANS, SDS AND INTERNAL CONSISTENCY RELIABILITY (ALPHA)

Form*	Case	Subtest	# Items	Mean (SD) Test 1	Mean (SD) Test 2	Reliability Test 1	Reliability Test 2
A	1	D	12	11.66 (0.64)	11.78 (0.49)	.25	.14
		P	21	17.29 (2.40)	17.28 (3.44)	.58	.80
		Case 1 Total	**33**	**28.94 (2.73)**	**29.06 (3.69)**	**.63**	**.80**
	2	D	12	11.46 (0.78)	11.41 (1.27)	.20	.72
		P	21	16.63 (2.87)	16.63 (4.10)	.81	.92
		Case 2 Total	**33**	**28.09 (3.32)**	**28.03 (4.86)**	**.81**	**.91**
	Combined	D	24	23.11 (1.28)	23.19 (1.62)	.53	.74
		P	42	33.91 (4.69)	33.91 (7.22)	.82	.93
		Test 1 Total	**66**	**57.03 (5.52)**	**57.09 (8.23)**	**.84**	**.93**
B	1	D	12	11.52 (0.91)	11.50 (0.83)	.49	.33
		P	21	17.82 (2.76)	18.21 (3.13)	.72	.81
		Case 1 Total	**33**	**29.33 (3.24)**	**29.71 (3.56)**	**.75**	**.81**
	2	D	12	11.52 (0.87)	11.50 (0.83)	.43	.36
		P	21	17.88 (3.69)	18.32 (2.88)	.89	.85
		Case 2 Total	**33**	**29.39 (4.21)**	**29.82 (3.24)**	**.89**	**.83**
	Combined	D	24	23.03 (1.61)	23.00 (1.58)	.68	.65
		P	42	35.70 (5.75)	36.53 (4.81)	.89	.85
		Test 2 Total	**66**	**58.73 (6.77)**	**59.53 (5.72)**	**.90**	**.87**

*There were 35 students receiving form A at the first testing and form B at the second testing. There were 33 students receiving the forms in reverse order.

TABLE 2

SAME TEST INTER-CASE CORRELATIONS

Form	Subscore	Raw Correlation Test 1	Raw Correlation Test 2	Corrected Correlation[1] Test 1	Corrected Correlation[1] Test 2
A	Descriptive	.62	.61	2.97	1.92
	Prescriptive	.58	.83	.85	.97
	Total	**.66**	**.85**	**.92**	**1.00**
B	Descriptive	.64	.82	1.39	2.38
	Prescriptive	.58	.28	.72	.33
	Total	**.65**	.41	**.80**	**.48**

[1]Raw correlations were corrected for the unreliability of their component scores by the standard disattenuation formula:

$$\text{Corrected Correlation} = \frac{r_{xy}}{\sqrt{r_{xx} \cdot r_{yy}}}$$

where

r_{xx} = reliability of the x score
r_{yy} = reliability of the y score
r_{xy} = raw correlation between x and y

TABLE 3

RAW (CORRECTED) INTERCORRELATIONS AMONG CASES FROM DIFFERENT TESTS[1]

I. Descriptive Subscore

	Test 2			
	Group 1		Group 2	
Case	1	2	1	2
Test 1				
1	.51 (1.78)	.38 (1.45)	.47 (1.81)	.78 (3.22)
2	.57 (1.72)	.56 (2.07)	.71 (1.19)	.71 (1.28)
Total		.58 (.99)		.82 (1.16)

II. Prescriptive Subscore

	Test 2			
	Group 1		Group 2	
Case	1	2	3	4
Test 1				
1	.58 (.85)	.64 (.79)	.69 (.91)	.76 (.90)
2	.48 (.68)	.37 (.45)	.79 (.98)	.72 (.80)
Total		.73 (87)		.87 (.96)

[1]Values in this table are interpreted as follows. Group 1 took form A at the first test administration and form B at the second. Thus, the first entry in the table (.51) represents the correlation of the descriptive subscore from the first case on form A with that of the first case on form B administered one week later. The value 1.78 was computed by the formula shown in the footnote to table 2. The .58 value shown under Total was computed by summing the prescriptive subscores over both cases administered at same time and correlating them.

REFERENCES

Chase, W.G., Simon, H.A. (1973). Perception in chess. Cognitive Psychology, 1, 55-81.

Elstein, A.S., Shulman, L.S., Sprafka, S.A. (1978). Medical problem solving: An analysis of clinical reasoning. Cambridge, MA: Harvard University Press.

Hively, W., Maxwell, G., Rabehl, G., Sension, D., Lundin, S. (1973). Domain-referenced curriculum evaluation: A technical handbook and a case study from the minnemast project. In M. C. Alkin (Ed.), CSE Monograph Series in Evaluation, 1. Los Angeles: Center for the Study of Evaluation, University of California.

Jacobs, R.M. (1988). Ten-year study of strategies for teaching clinical inference in predoctoral orthodontic education. Journal of Dental Education, 52(5), 235-244.

Norman, G.R., Feightner, J.W., Tugwell, P., Muzzin, L.J., Guyatt, G. (1983). The generalizability of measures of clinical problem-solving. Proceedings of the 22nd Annual Conference on Research in Medical Education. Washington, DC: Association of American Medical Colleges, 110-4.

McGuire, C., Babbott D. (1967). Simulation technique in the measurement of problem-solving skills. Journal of Educational Measurement, 4, 1-10.

Vu, N., Dawson-Saunders, B., Barrows, H. (1987). Use of a medical reasoning aptitude test to help predict performance in medical school. Journal of Medical Education, 62(4), 325-335.

Webster, G.D., Shea, J.A., Norcini, J.J., Grosso, L.J., Swanson, D.B. (1988). Strategies in comparison of methods for scoring patient management problems: Use of external criteria to validate scores. Evaluation & The Health Professions, 11(2), 231-248.