The Alberta Journal of Educational Research Vol. XLV, No. 3, Fall 1999, 278-287

John O. Anderson
University of Victoria

Modeling the Development of Student Assessment

The study reported in this article was an attempt to investigate how teachers evaluate the achievement of their students. The study was based on a dataset created from portfolios containing achievement products such as written assignments and tests and background information for a simulated student in a grade 8 language arts class. The contents of the portfolio were controlled in terms of achievement level of products and the background of the student. As part of an undergraduate teacher education course in classroom assessment, 147 teacher candidates graded the components of an assigned portfolio over a 10-week period and reported a final grade. These scores and grades were the basis for an investigation of the structure underlying the evaluation of achievement by these teacher candidates. The model developed fitted the data well and had resonance with commonsense views on the kinds of factors that could be expected to affect the decisions teachers make about marking student assignments and tests.

La recherche décrite dans le présent article constitue une tentative qui visait à étudier la façon dont les enseignants évaluent les réalisations de leurs élèves. L'étude repose sur une base de données créé à partir de dossiers renfermant diverses épreuves telles des travaux écrits et des examens, ainsi que des renseignements complémentaires sur un élève fictif en 8e année dans un cours des arts du langage. Le contenu des dossiers était contrôlé quant aux résultats obtenus aux épreuves et à l'information personnelle sur l'élève. Dans le cadre d'un cours de premier cycle sur l'évaluation en salle de classe, 147 étudiants en pédagogie ont évalué le contenu du dossier qu'on leur a remis. Le projet a duré 10 semaines et a pris fin quand les élèves ont accordé une note finale à l'élève fictif. Ces évaluations et ces notes finales forment la base d'une étude sur la structure sous-jacente à l'évaluation des réalisations telle qu'effectuée par des élèves en pédagogie. Le modèle développé convenait aux données et les résultats reflètent des conclusions de sens commun sur les facteurs qui pourraient influencer les enseignants dans leur évaluation des travaux et des examens.

The evaluation of student achievement is a characteristic and significant component of formal instruction. The completion of tests, assignments, projects, journals, and portfolios for evaluation purposes are typical student activities in the classroom teaching and learning environment (Stiggins, 1997; Ward & Murray-Ward, 1999). The assessment plan for a particular course or unit of instruction is generally described in terms of the tasks to be completed by the student as part of course requirements, the weighting to be assigned to each task, the marks or grades to be assigned (often including the mark to letter grade conversions) and a brief description of the meaning of each letter grade (e.g., an *A* is for *superior performance*, a *B* for *outstanding performance*). It is understood that a teacher will take examples of a student's work, mark them, and then combine these marks into a final grade. However, how a teacher

John Anderson teaches measurement and evaluation in the Department of Educational Psychology and Leadership Studies, of which he is also the Chair.

translates student products such as essays and test responses into marks or grades is not well researched (Brookhart, 1993; Cross & Frary, 1999).

The study reported in this article was an attempt to investigate how teachers evaluate the achievement of their students. The study was based on a dataset developed by Wilson and Shulha of Queen's University (Wilson, 1996). They created a set of portfolios containing achievement products (such as written assignments and tests) and background information for a simulated student called Chris in a grade 8 language arts class. The contents of the portfolio were controlled in terms of achievement level of products and the background of the student, resulting in a number of Chrises. As part of an undergraduate teacher education course in classroom assessment, 147 teacher candidates graded the components of an assigned portfolio over a 10-week period and reported a final grade for Chris at the end of the term. These scores and grades were the basis for an investigation of the structure underlying the evaluation of achievement by these teacher candidates.

The Data

The data consisted of the scores and grades generated by the 147 teacher candidates for the components of a student portfolio they were given over a 10-week period. The components were added to each portfolio on a weekly basis over this period. A total of eight student achievement products were included in the portfolio in addition to information about the background of the student. Most of the products were presented in bundles of three, and all three were included in the portfolio; one was identified as belonging to Chris. The achievement products included in the portfolio were:

1. "A Trip to the Mall": A written piece that had a maximum score of 25. This assignment had three levels.
2. "Salmon for Simon": A multiple-choice item test of reading comprehension that had a maximum score of 9. This assignment had three levels.
3. "Did I Order an Elephant?" A Cloze-format test of reading comprehension with a maximum score of 15. This assignment had three levels.
4. "New Kid on the Block": A short-answer, open-ended format assignment on reading that had a maximum score of 18. This assignment had three levels.
5. "Ghost Ship of Mahone Bay": A multiple-choice format test of reading with a maximum score of 9. This assignment had three levels.
6. "Mending Wall." This was a writing and editing assignment completed on the computer to represent a student's best work with a maximum score of 25. This assignment had three levels.
7. "School Dance": A written piece with a maximum score of 25. This assignment had a single level.
8. Final Examination. This was a mixed-format test (matching, identification, multiple-choice, and short-answer) with a maximum score of 130. This assignment had a single level.

Background information on Chris was presented indirectly in the portfolio in the form of notes, memos, and school reports:

1. *Expectations.* Expectations were to be inferred from information on student scores on the Canadian Tests of Basic Skills, parental occupational status, a

student essay entitled "Meet Me," and an accommodation report. There were three levels of expectation: low, medium, and high.

2. *Parental involvement*. Parental involvement was to be inferred from school memos and notes related to parental involvement with school activities such as parent interviews or volunteer work. There were two levels of involvement: low or high.
3. *Gender*. Chris's gender could be inferred from an audiotape of Chris reading a passage of text for miscue analysis.
4. *Growth*. Growth was present in three levels: *falling behind*, performing *steadily*, or *improving*. The three levels of growth were realized in two ways in the portfolio: the quality of student achievement products and in some ancillary information about Chris's performance in other school subjects as evidenced by reports from other school areas such as mathematics, science, history, and Resource Centre that were included in the portfolio. The quality of student achievement products varied according to the level of growth assigned that portfolio, the pattern of inclusion of materials followed the design for incorporating growth into the portfolios.

Table 2 in Wilson and Martinussen (in press), which precedes this article in this journal issue, summarizes the descriptive statistics for the scores awarded the different assignments and tests and the final grade on the report card. The correlations between scores (Wilson & Martinussen, in press, Table 4) show generally positive, low to moderate linear relationships between scores. There appears to be no single achievement product dominating the final grade (Report Card) for Chris. However the task "Salmon for Simon" has negative correlations with three other products ("Trip to the Mall," "Mending Wall," and "Ghost Ship of Mahone Bay") and a near-zero relationship (r=–.07) to the final grade on the report card.

Development of a Model

The main factor underlying the evaluation of the assignments and tests completed by Chris was thought to be *achievement* (Figure 1): the knowledge and skills Chris used to write passages, respond to questions, and complete tests and an examination. This conceptualization is compatible with the dominant view of educational measurement that testing is essentially unidimensional (Gulliksen, 1950; Lord, 1980) in that student achievement in language arts, for example, is based on a single underlying trait. The unidimensional construct would mean that a student with higher achievement, for example, should write a better essay than a student of lower achievement. The better essay should obtain a higher score than an inferior essay regardless of who the student is, whatever background he or she comes from, or whatever past performances (achievements) the student has attained. Achievement was not directly measured and so was viewed as a latent variable and is represented as an oval in the model described in Figure 1. It was thought that, despite the theoretical measurement perspective of unidimensionality, student background could enter into decisions about student achievement, and so this was also modeled as a latent variable. Another possible influence on marks could be perceived growth of students over time; this was structured in the portfolio by means of

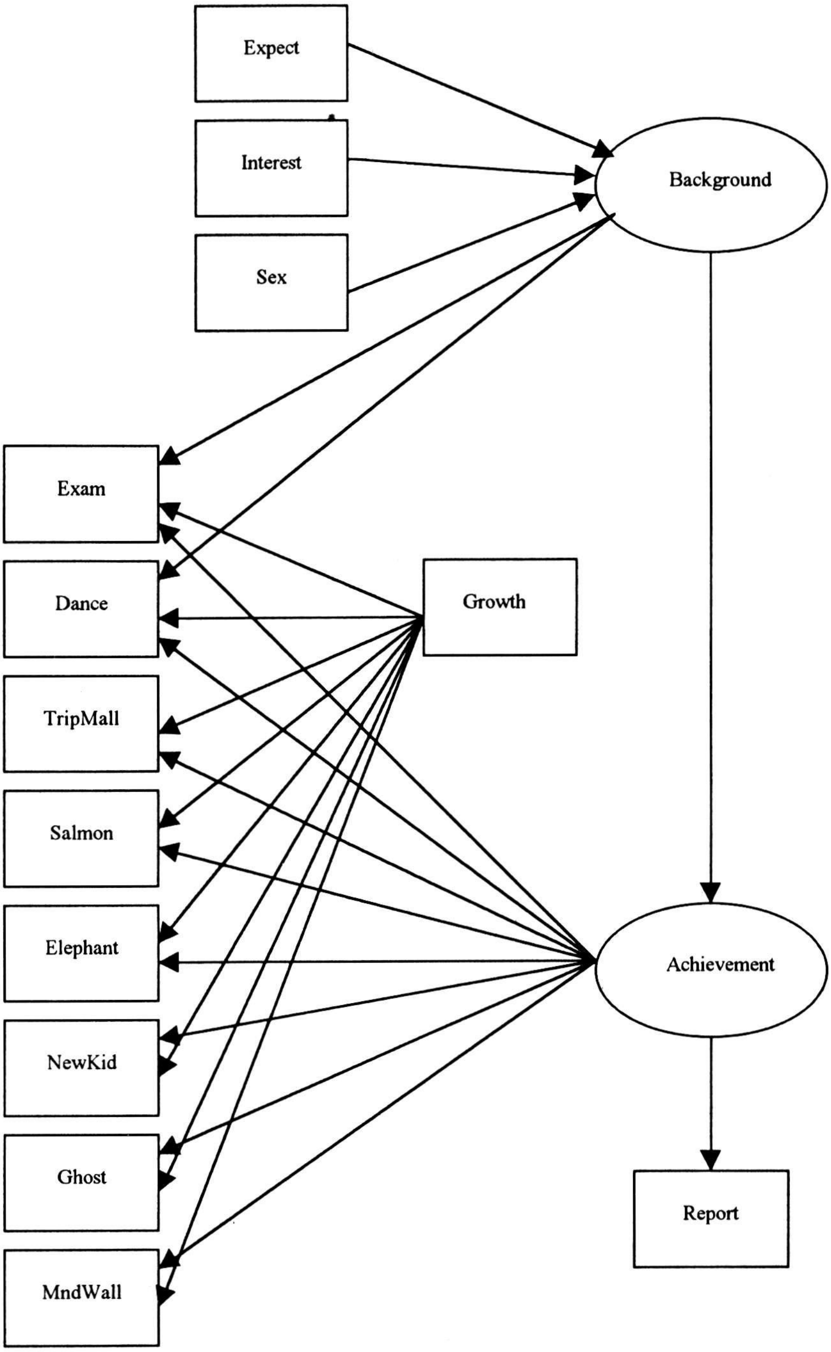

Figure 1. Model of achievement, growth, and student background (n=147, χ^2=47.4, p=.58).

notes and memos. Measured or observed variables were the eight achievement products, the background variables, and growth, and are represented as boxes.

It was of particular interest that both Final Examination and "School Dance," although the same products in every portfolio, showed substantial variance in the marks awarded and a positive correlation to the final grade (Report Card). If achievement were the only factor underlying the scoring and grading of Chris, the scores awarded Final Exam and "School Dance" should show little variation relative to the other achievement products across the 147 teacher candidates and should have a low correlation with the final grade awarded Chris. However, the summary statistics suggest that these two achievement products generated scores that were similar to those of other achievement products that did vary from one portfolio to the next. The interest focused on the question: Was there some source of this variation or prediction of the scores in the dataset? The attention fell on the background variables of *expectations, parent interest,* and *gender* as possible predictors of the scores awarded the Final Examination and the written assignment "School Dance."

In studies on assessment practices in the classroom (Bachor & Anderson, 1994) teachers reported the phenomenon of *gut feeling* in which the teacher would somehow develop a global estimate of the performance or achievement level of students in the class and all assessment results for an individual student would be related to this global estimate. Perhaps the above-mentioned background variables serve as a source for the development of this global achievement estimate that conditions the evaluation of particular achievement products by the teacher. So if this holds, a student who is considered a high achiever will tend to be awarded high marks, and a low achiever is awarded low marks for identical performance on an achievement task such as a final examination or a written assignment. In other words, factors other than achievement underlie the scores and grades teachers generate. One such factor hypothesized was background, which was modeled as a composite latent variable comprising the measured variables expectations, parent interest, and gender. It should be noted that the latent variable background has been modeled as being influenced by the three measured variables, and the direction of the arrow reflects this. This is somewhat different than usual in that the latent variables or factors in a model most often influence the measured variables (Loehlin, 1987; MacCallum, 1995), but this model does represent the design of the portfolio in that the three variables were explicitly included as indicators of student background. As noted, both Final Examination and "School Dance" were exactly the same in each portfolio and thus were not viewed as being influenced by level of student achievement; rather, both modeled to be directly influenced by background. However, it was considered reasonable that background of student may influence achievement and this is reflected in the model developed.

Another influence on marks awarded to achievement products was thought to be growth, a variable embedded in the portfolio in two ways: Chris's work contained in the portfolio showed growth, decline, or stability over the 10-week period of the evaluation; and reports that were included in the portfolio that showed Chris's progress in other school areas such as mathematics, history, and science. There were three levels of growth: *improving, steady,* and *falling* *behind*. All achievement products were modeled as being directly influenced by growth.

The final model, then, had three components influencing the evaluation of student achievement products: background, growth, and achievement. The final global evaluative description of Chris was the percentage grade awarded on the Report Card, and this was modeled as being influenced solely through achievement.

Analysis

The analysis of the data was conducted with AMOS 3.1 (Arbuckle, 1993), a program that allows for structural equation modeling. The dataset used in the structural equation modeling was slightly modified to account for missing data. The missing data were estimated by the mean values for the variable. There were no more than four missing values for any of the six variables with missing data. The model depicted in Figure 1 was submitted for analysis and resulted in a χ^2 of 47.4 (p=.58), which suggests that the data fit the model. The standardized regression weights are listed in Table 1. Squared multiple correlations are listed in Table 2.

Given the nonsignificant fit statistic for the model submitted for analysis (Figure 1), it can be concluded that the model fits the data well. Further, for the individual elements in the model the standardized regression of achievement on Report Card (.997) and the squared multiple correlation of Report Card (.990) suggest that the final grade awarded Chris is well accounted for by the model. However, there are interesting and potentially significant patterns of relationships among the achievement products and the components that are modeled as influencing the scores awarded.

All achievement products with the exception of Final Examination have significant regression weights and squared multiple correlations in excess of .10, suggesting that the data associated with the achievement products fit in the model. However, the results also suggest that different products are influenced by different factors. For example, "Trip to the Mall," "New Kid," and "School Dance" appear to be more influenced by achievement, whereas "Ghost Ship" and "Mending Wall" seem to be influenced primarily by growth. Also, the direction of influence varies by achievement product and underlying component. Achievement has positive effects on all achievement products, yet growth has positive effects on some and negative effects on others. As an example of negative effect, "Salmon for Simon" has a negative regression weight (standardized) with growth (–.756) indicating that with all other variables held constant, a 1-standard deviation increase in growth will be associated with a drop of .756 standard deviation in the score awarded "Salmon for Simon," whereas for "Ghost Ship" the score would increase .857 of a standard deviation. An expectation was that relationships between the marks awarded different achievement products would have consistent and positive correlations, particularly if achievement were the main underlying factor.

The two fixed achievement products have different relationships in the model. "School Dance" appears to be well accounted for in the model, with an R^2 of .31 and significant regression weights with all three underlying components, although a negative weight with growth. On the other hand, Final

Table 1
Standardized Regression Weights

Path			*Regression Weight*	*Significance*
Expectation	→	Background	.946	sig
Parent Interest	→	Background	.311	ns
Gender	→	Background	.093	ns
Background	→	Achievement	.214	ns
Achievement	→	Report	.997	sig
Achievement	→	Trip to Mall	.478	sig
Achievement	→	Salmon	.137	sig
Achievement	→	Elephant	.311	sig
Achievement	→	New Kid	.434	sig
Achievement	→	Ghost Ship	.170	sig
Achievement	→	Mending Wall	.360	sig
Achievement	→	Final Exam	.178	ns
Achievement	→	School Dance	.391	sig
Growth	→	Final Exam	−.064	ns
Growth	→	Trip to Mall	.199	sig
Growth	→	Salmon	−.756	sig
Growth	→	Elephant	−.225	sig
Growth	→	New Kid	−.114	ns
Growth	→	Ghost Ship	.857	sig
Growth	→	Mending Wall	.547	sig
Growth	→	School Dance	−.275	sig
Background	→	Final Exam	−.084	ns
Background	→	School Dance	.219	sig

Examination is not well accounted for by the model with an R^2 of .04 and has no significant regression weights with any of the underlying components in the model suggesting that much of the variance in the scores awarded by the teacher candidates for this achievement product is not accounted for by the three underlying components of background, growth, and achievement, but by factors not included in the model. What these factors may be could only be based on speculation at this time, although they would probably include the teacher's attitudes toward the purpose of student evaluation, the teacher's knowledge of language arts and the individual approaches used in grading student work.

Another focus of interest in regard to the adequacy of the model is the presence of correlated residual error (Table 3). These correlations could be taken to mean that some other factor accounted for variance in the scores of "Trip to the Mall," "Salmon for Simon," "New Kid," and "Ghost Ship." However, given the dataset used in this analysis, specification of this underlying influence could not be pursued. In analyzing these data several alternate models were tested but did not improve overall fit, nor were additional variables significantly related to student achievement results. For example, the gender of the teacher candidates was not significantly related to the grades

Table 2
Squared Multiple Correlations

Variable	R^2
Background	.99
School Dance	.31
Final Exam	.04
Trip to Mall	.27
Salmon	.59
Elephant	.15
New Kid	.20
Ghost Ship	.76
Mending Wall	.43
Report Card	.99

awarded, nor were the particular instructors of the teacher candidates in their undergraduate assessment course.

Discussion

This study was an exploratory inquiry of the use of student portfolios for the investigation of teachers evaluating student achievement. The model developed and investigated did fit the data well and had resonance with commonsense views on what kinds of factors affect the decisions teachers make about marking student assignments and tests. From a technical measurement perspective, it could be argued that only achievement should be the basis of the mark if validity of measurement is to be maximized, assuming the assessment tasks are valid representations of curricular intent. However, other research (Bachor & Anderson, 1994; Cross & Frary, 1999) has shown that teachers do use background information and general perspectives about students to condition the marks they give to students. This and experience as both a student and a teacher would lead one to expect that the components of achievement, background, and growth would all be elements teachers factor into the mark given a student.

The analysis presented in this study shows that the assessment of student achievement is not a simple unidimensional phenomenon, but rather a complex process that involves a number of underlying factors that vary in influence

Table 3
Correlations of Residual Error Terms

Error Terms			*Correlation*
Trip to Mall	↔	School Dance	.22
Salmon	↔	School Dance	–.26
Trip to Mall	↔	Salmon	–.23
Salmon	↔	Elephant	.21
Trip to Mall	↔	Ghost Ship	–.21
Ghost Ship	↔	School Dance	–.11

from one achievement product to the next. To assume that a single factor such as achievement underlies the marks awarded would be a mistake, although from a measurement perspective it would be the desired situation. In fact with these data, growth appears to have more influence over the marks awarded achievement products than does achievement. Yet the influence of an underlying factor does not have a consistent influence across all achievement products. Growth, for example, has a positive influence on "Ghost Ship" and a negative influence on "Salmon for Simon," both multiple-choice format tasks in language arts. This difference in effects cannot be easily explained with these data, yet is a significant issue for future research.

The variation in marks awarded to the common products "School Dance" and Final Examination is troubling in a couple of ways. First, given that both of the products were identical in all of the 147 portfolios, the mark that a student is awarded depends largely on which teacher marks the assignment or exam. Second, the variation of the written assignment "School Dance" is relatively well accounted for by the model, suggesting that perhaps the marks awarded can rely on an achievement gestalt developed by the teacher and consisting of an amalgam of background information and perceptions of the student that is independent of the actual performance of the student on achievement-relevant tasks. This issue warrants further investigation because of the seriousness of its consequences.

In conclusion, one of the major outcomes of this study is the demonstration of the utility of this approach of controlled portfolios of student achievement products and background information, and structural equation modeling, for investigating complex phenomena such as the assessment of student achievement. The use of the approach points out significant areas for future research activities. These include the further development of structured portfolios as a research tool for studying the assessment of the student achievement: this approach has the potential for enhancing the understanding of this largely unexplored yet pervasive element of education. The model that underlies the decisions educators use to evaluate student achievement is not likely to consist of a single factor, but would be more complex. In using the portfolio approach to investigate this area, more information structures should be incorporated into the student dataset such as more and deeper personal information (e.g., journal writings and perhaps student photos), indicators of attitudes and feelings, more varied achievement products of a formative nature, and more information about the instructional context. In a continuation of this study the teacher candidates have been requested to keep a journal of their thoughts and feelings as they completed the tasks associated with evaluating student achievement. The analysis of these data along with their marks and grades should shed further light on the structures underlying the evaluation of student achievement.

References

Arbuckle, J. (1993). *AMOS: Analysis of moment structures 3.1.* Philadelphia, PA: Temple University.

Bachor, D.G., & Anderson, J.O. (1994). Elementary teachers' assessment practices as observed in the province of British Columbia. *Assessment in Education, 1*(1), 65-95.

Brookhart, S. (1993). Teachers' grading practices: Meanings and values. *Journal of Educational Measurement, 30,* 123-142.

Cross, L.H., & Frary, R.B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education, 12*(1), 53-72.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Loehlin, J.C. (1987). *Latent variable models.* Hillsdale, NJ: Erlbaum.

Lord, F.M. (1980). *Applications of item response theory to practical test problems.* Hillsdale, NJ: Erlbaum.

MacCallum, R.C. (1995). Model specification: Procedures, strategies and related issues. In R.H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications.* Thousand Oaks, CA: Sage.

Stiggins, R.J. (1997). *Student-centered classroom assessment* (2nd ed.). Upper Saddle River, NJ: Merrill.

Ward, A.W., & Murray-Ward, M. (1999). *Assessment in the classroom.* Belmont, CA: Wadsworth.

Wilson, R.J. (1996, June). *Factors affecting the assessment of student achievement.* Paper presented at the annual meeting of the Canadian Educational Researchers' Association, Brock University, St. Catharines.

Wilson, R.J., & Martinussen, R.L. (in press). Factors affecting the assessment of student achievement. *Alberta Journal of Educational Research.*