Visual Testing: Searching for Guidelines

Kitty Van Gendt
University of Groningen, Netherlands

Pløn Verhagen
University of Twente, Netherlands

Abstract

An experiment was conducted to investigate the influence of the variables 'realism' and 'context' on the performance of biology students on a visual test about the anatomy of a rat. The instruction was primarily visual with additional verbal information like Latin names and practical information about the learning task: dissecting a rat to gain insight in the anatomy of a mammal. Students were tested on: (a) recognition of anatomical objects, (b) labeling of these objects and (c) relations between objects. Results indicate that the amount of realism and context used in the text influences test performance depending on the learning tasks. Test results also show a learning hierarchy in the different learning tasks with the recognition task being the easiest and the relations task the most difficult.

Introduction and the Context of the Problem

In several courses that are taught in the faculty of Biology at the State University of Groningen in the Netherlands, it is considered necessary that students learn about the anatomy of a mammal by dissecting the body of a rat. The students perform this task in a two-day practical in which they dissect dead animal material (the rat) layer by layer to identify the different anatomical structures such as the different muscles and organs. This practical will always exist in this way as the faculty staff takes the firm position that the students need the experience of the stepwise dissection of a real rat to gain thorough understanding of these anatomical structures. Careful dissection of the animal enables the student to examine every part of the rat in detail and to get a grip on what they have to learn.

The problem occurred due to the teaching approach in which the students had to report about their observations by making anatomical drawings of the anatomical structures according to drawing rules that were part of the instructions of the course. The drawings of the anatomical structures had to depict all parts involved with their relative proportions and their interconnections, together with the Latin names of the parts in a legend in the margin of the drawing. After two days work, each student had produced a set of drawings of a rat. The problem was that a limited number of staff members had to correct the work of large numbers of students. There are about 150 students yearly, distributed over two courses for first-year biology students and second-year pharmacy students, who produced about ten drawings each. The correction of these drawings is so time consuming, that the students sometimes had to wait for months before they get feedback on their performance. This situation was educationally undesirably. Further the variety of drawing skills of the students introduced a subjective element in accepting or rejecting a drawing as a correct presentation of a given anatomical structure. Moreover, different members of the teaching team put different accents while rating student performance. Uniform evaluation of the drawings could therefore not be guaranteed. This situation was highly unsatisfactory. So the problem was that the variety of the students drawing skills leads to a subjective judgment. Besides this the spread of the correction work over several teachers caused a non-uniform evaluation and last the students received delayed feedback.

These limitations resulted in the need of the teaching staff to look for other possibilities for assessment. That need motivated this research project. The identified assessment problem is, however, more general than the problem with the drawings in the particular practical about the anatomy of the rat. The learning outcomes that have to be assessed are largely in the visual domain (the anatomical parts and their relative positions and interconnections), which led to the choice to study the assessment problem on a more general level in the field of visual testing. The concrete problem of the practical about the anatomy of the rat will then be used as the test bed for that study.

Visual learning - Visual testing

In searching for alternatives for the assessment on the basis of student drawings, the construct to be learned during the instruction had to be redefined. According to Cronbach and Meehl (1955) "a construct is some postulated attribute of people that is assumed to be reflected in test performance. In test validation the attribute about which we make statements in interpreting a test is a construct". In the case of the anatomy of the rat the construct to be learned consists of (a) the decisive visual characteristics of anatomical objects like shape, size, texture and color of the object, (b) the spatial relationships between the objects and the relative size of each object in relation to other objects, (c) the verbal labels and descriptions of the objects, and (d) the functional relationships of objects. This construct is essentially visual, whereby the verbal labels and descriptions extend the qualities of the construct into the verbal domain, allowing for verbal communication about the spatial structure and its components

The definition of the learning construct resulted in the definition of learning tasks for the instruction. After following the course, the students should be able to (a) recognize the different anatomical structures, (b) label each structure by its Latin name, (c) specify their spatial orientation and (d) specify their functional interconnections. The learning outcomes can subsequently be assessed by requiring the students to demonstrate that they are able to fulfill three tasks: (a) the recognition task, (b) the labeling task, and (c) the relations task. In the recognition task the students must recognize the correct object according to its visual characteristics. In the labeling tasks the students has to identify an object by its Latin name. The relations task is a task in which the student has to show understanding of the spatial relations between objects by their functionality.

Although the instruction is primarily visual, it is relevant to know whether the assessment should also be visual. According to the stimulus generalization theory (Hartman, 1961) learning increases as the testing mode approximates the mode in which the information was presented. Dwyer (1978) gave an example in which he stated that instruction presented via a visual modality but evaluated in a conventional pencil-and-paper assessment would probably not provide an accurate representation of the total amount of learning that has occurred. The effectiveness of instruction presented to the students through a visual channel might most appropriately be measured by employing criterion measures assessing contributions of the visual mode of instruction.

As in all learning also in visual learning the more complex contructs contain concepts and relationships that are hierarchically ordered. Smith and Ragan (1999) summarized theories about learning hierarchies which all prescribe a similar pattern: learners first should be able to recognize a concept in order to identify that concept by its name. The acquired concepts then form prerequisites for more difficult tasks such as problem solving. Dwyer (1978) made a distinction in phases in a learning hierarchy for visual learning, starting with facts and definitions in a content area that are familiar to a person. In this way the person is prepared to relate and combine known and new elements to form new concepts. The more concepts a person possesses, the easier it is to form generalizations and rules. These processes are again prerequisites for problem solving. Dwyer (1978) states that: "the implication to be derived from the concept of a learning hierarchy is that since there are different kinds of educational objectives there also are different kinds of learning, each requiring students to perform different kinds of activities and each possessing unique conditions for optimum learning to occur". This assumes that in designing a visual test, it is required to look more closely to the type of item format that will assess these different kinds of learning in a valid way. In this case, the recognition task of the visual concepts is a prerequisite for identifying that concept with the correct name. Together they are prerequisites for the relations task.

Item types

For visual testing, Dwyer (1978) designed the PSE-test (PSE: Program of Standardized Evaluation) in which he tested the learning effect of instruction given verbally with additional visualization versus verbally alone. The subject of instruction is the anatomy of the heart. The test consisted of four parts: (a) a terminology test, (b) an identification test, (c) a drawing test and (d) a comprehension test. The terminology test measures the students knowledge of specific facts, terms and definitions. The identification test measures the students ability to identify parts or positions of an object. The drawing test measures the students ability to construct and/ or reproduce items in their appropriate context and the comprehension test measures the students understanding of the heart, its parts and its internal functioning. These four tests combined form an overall test for measuring the student's total understanding of all the content material. The identification test, terminology test and comprehension test consisted of multiple choice items only. The test items for the comprehension test was the most difficult since they were designed to measure the student's understanding of complex procedures and processes.

The use of multiple choice items is sometimes being criticized. Martinez (1999), for instance, who is also active in the field of visual testing, claims that multiple choice items often elicit low-level cognitive processing. He designed constructive figural response items (CFR) which would evoke complex thinking and therefore be more appropriate for testing the student's understanding of complex procedures and processes (Martinez, 1994). These items differ from traditional items in two ways: (a) they require mental construction of a response, rather then selection among options, and (b) they require demonstration of proficiency in a figural medium. Martinez (1990) argued that comparison of multiple choice items with their figural constructive response counterparts showed that CFR items were more difficult, more discriminating and more reliable. Martinez (1999) argues that the use of CFR items is best for items that evoke complex thinking. Other research outcomes (Martinez & Jenkins, 1993) were that CFR items were better able to distinguish between novices and experts. CFR items are sometimes referred to as free response or open-ended items.

Martinez (1993) and Parshall, Davey and Pashley (2000) recognized the fact that as technology improves it becomes more useful in visual testing because it gives possibilities to innovate item formats for visual testing by using objects and media in item formats. This may bring an interactive aspect in the item, for instance by requiring students who take the test to scale object size by dragging with the mouse or to move objects to required positions.

Variables for the study

The visual aspects of the construct to be learned are decisive visual characteristics like shape, size, texture and color of the objects. The representation of these characteristics in visuals is affected by the amount of realism of those visuals. 'Realism' is mostly associated with photographic pictures. More schematic representations such as line drawings are regarded to be less realistic, although they may be more effective to articulate certain visual characteristics than realistic pictures do. Dwyer (1978) claims that the effectiveness of realism in this sense depends on the learning tasks and the instructional method. According to Mandler and Ritchey (1977, in: Anderson, 1994), however, people are better in remembering the meaning of a visual then the details of that particular visual. Which would here mean that realism would not add to visual testing because the details are not

important since people only remember the meaning. This uncertainty was reason to choose 'realism' as the first variable for this study.

The construct to be learned also contains the recognition of objects with their spatial positions between other objects and the relative size of the objects in relation to each other. The availability of the other objects appears to influence the recognition of objects. Cave and Kosslyn (1993) conducted an experiment in which they speak of 'holistic pictures' when objects are embedded in a visual context. They found that the recognition of objects in holistic pictures resulted in higher mean scores then the scores for isolated objects. Research in the field of face recognition (Tanaka & Farah, 1993) resulted in a similar outcome. Tanaka and Farah demonstrated that recognition of facial components was facilitated by the presence of the facial contour. Isolated facial objects were difficult, in most cases impossible, to recognize correctly.

The results of Mandler and Ritchey as well as those of Tanaka and Farah suggest that 'visual context' is a relevant variable for visual testing. 'Context' was thus selected as the second variable for this study.

Hypotheses

The effect of the variables 'Realism' and 'Context is studied on the basis of three hypotheses. The contradictory results of earlier research about the amount of realism led tot the first hypothesis:

H1: There will be no difference in mean score between items with realistic color pictures and items with schematic drawings.

Hypotheses of this kind have mostly been studied with types of instruction that are primarily verbal with additional visuals. In our case, the instruction is primarily visual.

The variable 'Context' is studied by testing the second hypothesis:

H2: Items with contextual information will result in higher mean scores then items without contextual information.

Hypothesis three is based on the theory of learning hierarchy. The requirement that concepts of anatomical structures have to be learned as a prerequisites for understanding functional relations, lead to the assumption that this letter task is more difficult:

H3: Items testing the relations task will be more difficult then items testing recognition and labeling.

Item formats

Item formats were chosen based on the insights of Martinez (1999) and Martinez and Jenkins (1993) that recognition could be assessed with multiple choice items whereas items requiring more complex tasks such as relating objects based on functional relationships should be tested with CFR items. For the purpose of the study eventually the following item formats were used:

1. Multiple choice items
2. Labeling items
3. Connect-the-dots items

The multiple choice items all contained four alternatives in which the correct object had to be recognized. In the labeling items the examinee had to identify an object by its Latin name which could be chosen from a long list in an index. The connect-the-dots items are examples of CFR items. With the connect-the-dots items, examinees had to make connections between anatomical objects and also give the direction of these connections according to their functionality.

Experimental Design

Figure 1 gives an overview of the 2*2-experimental design in which there were two independent variables being the amount of realism and the context information. For the variable 'Realism' the choice was made to look at the influence of colorful realistic pictures (photo's) of anatomical structures versus the influence of schematic drawings of the same structures. The variable 'Context' was operationalized by looking at the influence of the presence of contextual information versus the absence of context. Figure 1 summarizes the experimental design.

		Realism	
		Picture	Drawing
Context	Yes	Picture + context	Drawing + context
	No	Picture - context	Drawing - context

Figure 1. Independent variables and conditions

De dependent variables are test scores for each of the three learning tasks of the course and a total test score for all test items together:

1. The mean test score on recognition items
2. The mean test score on labeling items
3. The mean test score on relation items
4. The mean test score on the overall test

Method and Procedure

The sample population consisted of 134 first year biology students. The students followed the dissection course for two full days. During these two days the rat was dissected in ten layers. The students were working in groups of thirteen and every student got an own rat for dissection. Before the course each student received a study guide with the instructional objectives and the dissection steps explained according to a cued instruction. The cues that were used in the guide were all Latin names of the anatomical objects that they had to learn. The cued instruction for each dissection layer followed a certain procedure in which the students as a group got a live demonstration of the dissection of that particular layer by a teaching-assistant. The teaching assistant used the Latin names to identify the different anatomical objects, show the important visual characteristics and explain the different functions of each object in accordance to their interconnections. After the demonstration, the students were required to dissect the current layer individually and accordingly observe the anatomical objects. Students had the freedom to take notes during the observation by making drawings of their observations. After completion of the individual dissection tasks, the students were asked to join the group and have a group discussion about the observation with the help of a poster that contained the outline drawing of the dissected layer. Every student had to label at least one anatomical object with its Latin name. At the end of the course on the second day, the student performed a computerized test. The test contained 57 items of which 19 items were based on recognition tasks, 20 items were based on labeling tasks and 18 items were based on relations tasks. The test was specifically developed for the purposes of the experiment. Three groups of questions were designed to measure achievement on the three different learning tasks: (a) multiple choice items, (b) labeling items, and (c) connect-the-dots items. For the picture-version (the realistic version of the variable 'Realism') photographs were taken of all needed visual materials. Image-processing software was used to produce the correct alternative and also the distracters for the multiple choice items, for isolating objects for the no-context-condition, and to add a graphical layer with dots to be connected for the connect-the-dots items. For the drawing-version of the test students of an art school were recruited to make the drawings. The testing program itself was programmed in C++.

Students were randomly assigned to one of the four conditions being: (a) picture with context, (b) picture without context, (c) schematic drawing with context and, (d) schematic drawings without context. At the start of the test, each candidate got an example of each item format to get acquainted with the style of questioning. The results of the test were analyzed quantitatively in SPSS 10.0 by comparison of means, ANOVA, Univariate Analysis of Variance and reliability tests.

Results

Table 1 gives an overview of the results of the univariate analysis of variance on the variables 'realism' and 'context'. The table gives an general overview of the significant differences for every learning task and the overall test.

Table 1. Univariate Analysis of Variance on variables 'realism' and 'context'

	Recognition		Labeling		Relations		Overall test	
	F	Significance	F	Significance	F	Significance	F	Significance
Realism	11.266	.001	2.076	.152	6.727	.011	7.877	.006
Context	38.129	.000	.485	.487	1.762	.187	.811	.370
Realism * Context	36.218	.000	.170	.680	.049	.049	.128	.721

The results are discussed below by realism, context and item difficulty.

Realism

Figure 2 shows the mean scores of the examinees for the condition realism of drawing versus colorful pictures on the different learning tasks and on the overall test. These mean scores are on a scale of 0 to 10.

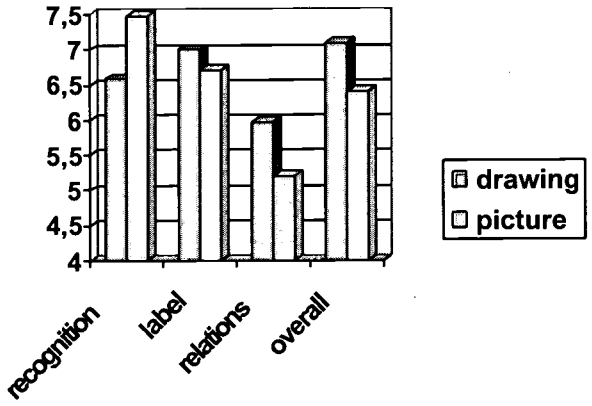

Figure 2. Results of the mean scores on the condition realism: picture vs. drawing,

As can be concluded from Figure 2 and Table 1, there were differences between drawing (for which the exact mean is M= 6.59) and picture (M= 7.48) on the recognition task in favor of picture with a significant difference (p = 0.001). For the labeling task there is a slight difference between the conditions in favor of drawing but this difference is not significant (p = .152). However, the difference found on the relations task between drawing (M= 5.98) and picture (M= 5.19) was significant (p = .011) in favor of drawing and for the overall test the difference between drawing (M= 7.10) and picture (M= 6.42) was also significant (p= .006) and in favor of drawing.

Context

Figure 3 gives an overview of the mean scores for the presence of context versus the absence of context on the different learning tasks and on the overall test. The scale of the mean scores is from 0 to 10.

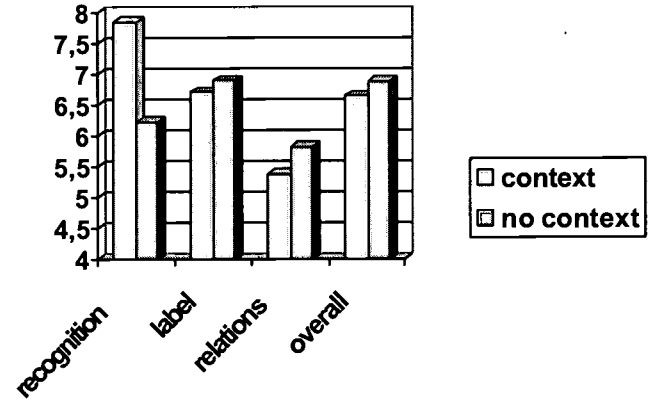

Figure 3. Results of mean scores on the condition context: with context vs. without context

As can be concluded from Figure 3 and Table 1, there were differences between the presence of contextual information in the test where examinees score higher for the recognition of objects (M= 7.84) then in the absence of contextual information (M= 6.21). This is the only significant difference (p = 0.00) that was found. The slight differences that were found for the remaining tasks were not significant which resulted in the overall test also showing no significant differences.

Realism * Context

An interaction effect was found between realism and context for picture (M= 7.48) and the existence of context (M= 7.84) with a level of significant of p= 0.00.

Item difficulty

Table 2 gives an over view on test reliability for Cronbach's alpha and the standard error of mean.

Table 2. Test reliability with Cronbach's alpha and Standard Error of Mean

	Ss	N (items)	Cronbach's alpha	Standard Error of Mean
Overall test	134	57	.8583	.066
Recognition items	134	19	.5873	.097
Labeling items	134	20	.7788	.061
Relations items	134	18	.7113	.097

Table 2 shows that the overall test is reliable with Cronbach's alpha being .8583 and an SEM of .066. The least reliable are the recognition items. Labeling and relations items are equally reliable. Figure 4 shows the item difficulty for the three learning tasks. The scale on the horizontal axis is from 0 to 10 and gives an overview of the proportions of students that answered a particular item correctly on the different learning tasks.

Figure 4 shows a difference in item difficulty with recognition tasks being the easiest and relations tasks the most difficult. Table 2 and Figure 4 complement each other. Figure 4 shows that low reliability of recognition items may be caused by a ceiling effect. The recognition items are obviously rather easy. The more difficult tasks are discriminating much better between learners who know and those who have learned less. The distribution of the items on the scale for the relations tasks shows that this task was the most difficult and the most discriminating. As this task was (thus) also very reliable, the results support the contention of Martinez (1990) who argues that these characteristics are typical for figural-constructive-response items. CFR items seem a proper choice for distinguishing between experts and novices.

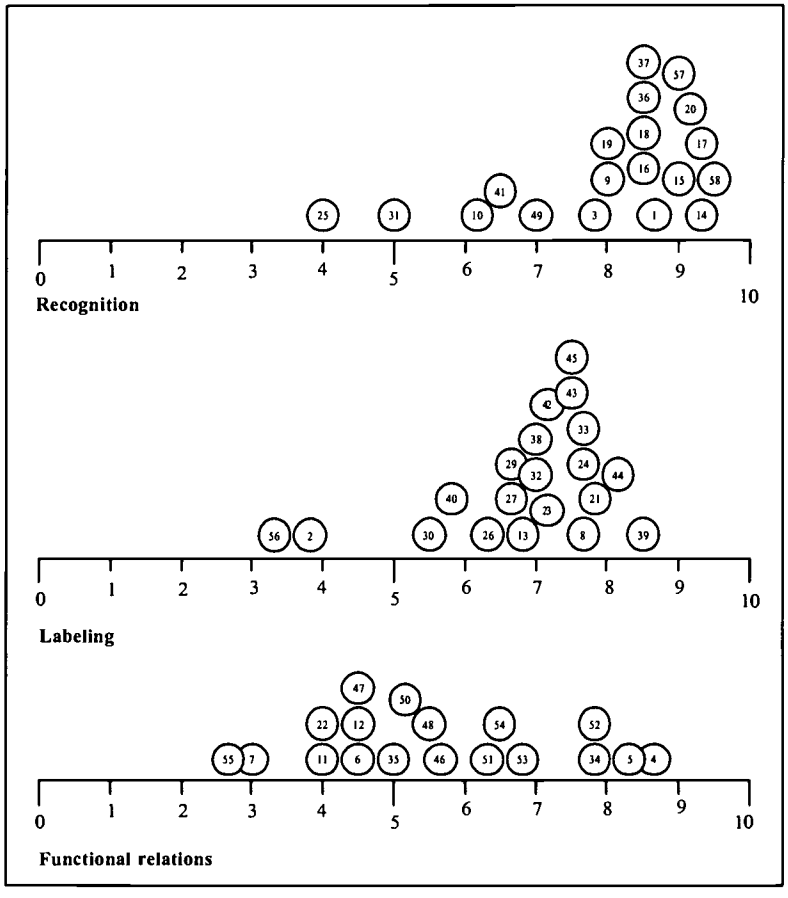

Figure 4: Item difficulty for each learning task

Discussion and Conclusions

Returning to the hypotheses stated earlier in the paper and the experimental results, some conclusions can be drawn on visual aspects of visual testing.

Hypothesis one

The first hypothesis states that we may expect no differences in mean scores between items with realistic color pictures and items with schematic drawings. This hypothesis must be rejected. We did find a significant difference in favor of picture over drawings for the recognition task. Also, a significant difference was found in favor of drawing over picture for the relation task and the overall test. These results suggest possible guidelines which are that for constructs that are primarily visual, it is best to test recognition of objects with pictures and to test relations between objects with drawings. The richer amount of stimuli in pictures thus seems to help recognition of necessary visual characteristics, while for the more abstract relations task, the main characteristics of the objects are sufficient. Drawings can be focused on these main characteristics. These findings comply with the results of Mandler and Ritchy (1977) who found that people are more sensitive to remember the meaning of a visual then the details.

Hypothesis two

The second hypothesis expects that items on contextual information will result in higher mean scores then items without contextual information. This expectation was only true for the recognition of objects. Thus it seems that for recognition tasks the examinees are helped by 'holistic visuals' in the sense of Cave and Kosslyn (1993) to succeed in answering the test items. For the other tasks the isolated objects gave sufficient information to complete the task.

Hypothesis three

The third hypothesis states that the relations tasks will be more difficult then the recognition and labeling tasks. This hypothesis cannot be rejected. The relation tasks appears to be the most difficult one. This finding supports the findings of Dwyer (1978) stating that recognition and labeling are prerequisites for relations. There is also the possibility of Martinez and Jenkins (1993) stating that CFR items are better able to distinguish between novices and experts then multiple choice items which is supported in our results in which the items on the relations task were the most difficult but had a wide spread in scores in contradiction to the items of the recognition and labeling tasks.

The main conclusions of the experiment are that:

1. Anatomical structures were better recognized as picture than as drawings.
2. Drawings yielded better results for test items on relations between anatomical structures than pictures.
3. The results for the overall test favor the use of drawings over pictures.
4. Context helped to recognize anatomical structures.
5. Scaling of item difficulty from easy to difficult shows the following order: Recognition - Labeling - Relations.

As far as guidelines for constructing visual tests are at stake, the results for 'Realism' and 'Context" gave reason for two guidelines:

1. The use of contextual information and color pictures can facilitate the recognition of objects.
2. In order to assess the knowledge of relations between objects based on spatial relationships or object functions, the use of schematic drawings can be sufficient.

In regard to item types, a guideline about CFR items seems to be justified:

1. For visual tests to distinguish between experts and novices constructive figural response items can be useful.

These guidelines need further study to get a more precise insight in the relationships between learning tasks and visual aspects. New technological possibilities for innovative item types, may extend definitions of visual constructs in ways yet to be found. The field of CFR items is thereby interesting as new assessment methods to measure particular constructs are there already technically available. Methods for scoring these items are a challenge and need further research for the development of suitable scoring algorithms. The current study was just a modest step to set the stage for further research on visual tests.

References

Anderson, J.R. (1994). *Learning and memory: an integrated approach.* NY: John Wiley & Sons Inc.

Cave, C.B., Kosslyn, S.M. (1993). The role of parts and spatial relations in object identification. *Perception, 22,* 229-248

Cronbach, L.J., Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52* (4) 281-301

Dwyer, F.M. (1978). *Strategies for improving visual learning.* State College, PA: Learning Services

Gagné, R.M. (1965). *The conditions of learning.* New York, NY: Holt, Rinehart and Winston.

Hartman, F.R. (1961) Recognition learning under multiple channel presentation and testing conditions. *AV Communication Review, 9,* 24-43

Martinez, M.E. (1990). A comparison of multiple choice and constructed figural response items. Princeton, NJ: ETS

Martinez, M.E. (1993). Item formats and mental abilities in biology assessment. *Journal of Computing in Mathematics and Science Teaching, 12* (3/4), 289-301

Martinez, M.E. & Jenkins, J.B. (1993). Figural response assessment: system development and pilot research in cell and molecular biology. Princeton, NJ: ETS.

Martinez, M.E. (1999). Cognition and the question of item format. *Educational Psychologist, 34* (4) 207-218.

Parshall, C.G., Davey, T., Pashley, P. (2000). Innovative item types for computerized testing. In: W.J. van der Linden, C.A.W Glas (Eds.), *Computerized adaptive testing: Theory and Practice,* pp.129-148. Dordrecht:Kluwer.

Smith, P.L., Ragan, T.J. (1999). Instructional design. 2nd Ed. New York, NY: Wiley.

Tanaka, H., Farah, M.J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Physiology. 46, A,* 225-246.