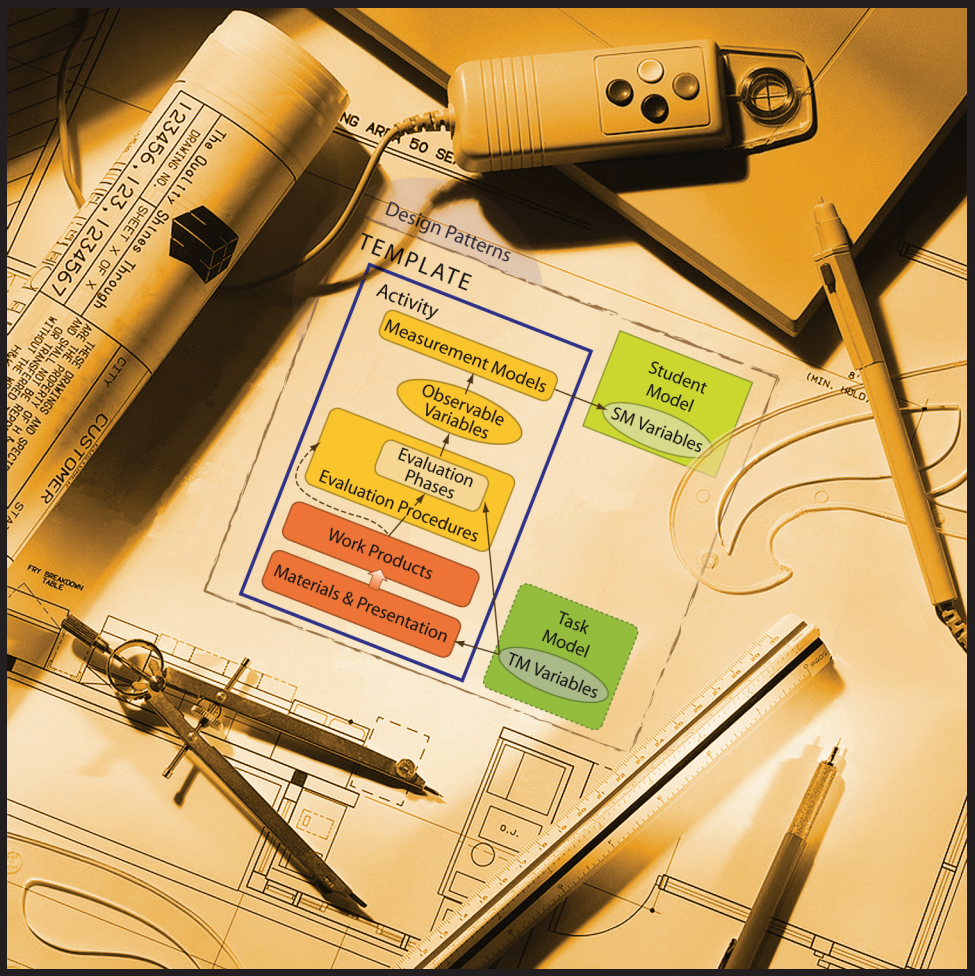



Large-Scale Assessment Technical Report 9| March 2010

# Informing Design Patterns Using Research on Item Writing Expertise

## Project: Application of Evidence-Centered Design to State Large-Scale Science Assessment

**Paul Nichols & Dennis Fulkerson**, Pearson

Report Series Published by SRI International

**SRI International**
**Center for Technology in Learning**
**333 Ravenswood Avenue**
**Menlo Park, CA 94025-3493**
**650.859.2000**
**http://ecd.sri.com**

**Technical Report Series Editors**
Alexis Mitman Colker, Ph.D., *Project Consultant*
Geneva D. Haertel, Ph.D., *Principal Investigator*
Robert Mislevy, Ph.D., *Co-Principal Investigator*
Ron Fried, *Documentation Designer*

APPLICATION OF EVIDENCE-CENTERED DESIGN TO STATE LARGE-SCALE SCIENCE ASSESSMENT

TECHNICAL REPORT 9

# Informing Design Patterns Using Research on Item Writing Expertise

**March 2010**

Prepared by:
Paul Nichols and Dennis Fulkerson
Pearson

*Acknowledgments*
This material is based on work supported by the National Science Foundation under grant DRL-0733172 (An Application of Evidence-Centered Design to State Large-Scale Science Assessment).

CONTENTS

# FIGURES

# LIST OF TABLES

## ABSTRACT

The development of a cognitive model of item writing expertise holds promise for informing the development of *design patterns*. Careful examination of item writers' processes and structures should provide insight into the design space specified by a *design pattern*. This technical report presents a study in which verbal reports from expert item writers were collected and analyzed. Findings from the study are used to suggest modifications to the development of *design patterns*.

# 1.0 PADI Design Patterns and a Cognitive Model of Item Writing Expertise

Educators are asking for large–scale assessment (LSA) tasks that assess higher level skills and are consonant with both instructional practice and learning science. Computers and other media offer potential solutions to the practical challenges posed in assessing complex constructs such as model-based reasoning and science inquiry in the context of challenging content. Computer–based assessment helps us assess student knowledge, skills, and abilities. Recent advances in technology allow test content to be delivered and presented to students in new ways, allow them to interact with tasks, and capture richer performance information.

Computer–delivered, technology–based tasks such as simulations and investigations that can address higher level skills and support learning have proved difficult and costly to develop for LSAs. Developing computer–delivered interactive tasks with procedures that evolved for multiple–choice items leads test developers to spend time "reinventing the wheel" (Riconscente, Mislevy, & Hamel, 2005). Current assessment research that intends to make the principles of assessment design explicit and to build conceptual and technological tools that can facilitate new assessment practices (NRC, 2001) holds great promise for developing computer–delivered interactive tasks.

The approach of evidence–centered assessment design (ECD) is an example of current assessment research that holds great promise for addressing the challenges of developing computer–delivered interactive tasks. This framework is both conceptual and technical. ECD provides a layered structure for assessment designers to explicate design rationale and its expression in the machinery of assessments. ECD represents assessment as an evidentiary argument through which we must reason from what we observe students say, do, or make in a few particular circumstances, to what they know, can do, or have accomplished as more broadly construed (Messick, 1994; Mislevy, et al., 2003).

The PADI project (http://padi.sri.com/) has developed broadly applicable data structures, representational forms, and software tools to support the application of ECD to developing computer–delivered interactive tasks. *Design patterns* are one class of tool developed under PADI to support the application of ECD to developing computer–delivered interactive tasks. A completed *design pattern* specifies a design space of elements to assemble into an assessment argument: Focal Knowledge and Skills, Characteristic and Variable Task Features, Potential Work Products, stimulus situations, and evaluation schemes. This design space focuses on the

science being assessed and guides the design of tasks with different forms and modes for different situations. *Design patterns*, in turn, ground *templates* for authoring more specific families of tasks. In the context of large-scale assessments, *design patterns* fill a crucial gap between broad content standards and particular tasks in a way that is more generative than test specifications and which addresses alignment through construction rather than retrospective sorting.

The development of a cognitive model of item writing expertise holds promise for informing the development of *design patterns*. Careful examination of item writers' processes and structures should provide insight into the design space specified by a *design pattern*. One example of how this improvement might take place is through incorporation into the *design pattern* additional information and processes gleaned from the examination of expert item writers' cognitive self-reports.

A cognitive model of item writing may be particularly helpful in informing the development of *design patterns* for innovative items. The use of innovative item types is increasing on large-scale assessments. Innovative items require student manipulation of graphic content in response to the item stem. One type of innovative item is figural response. Figural response items are novel test items useful in the assessment of science proficiency in computer–based test environments (Martinez, 1993). Due to the novelty of this item type, the development of high quality figural response items can be a challenging and difficult task for item writers. Consequently, item writers frequently produce poor quality or mediocre figural response items that require laborious refinement by professional test developers. There is a practical need to identify and describe the conceptual knowledge and skills that item writers need to possess to improve their figural response item writing.

The research reported here takes the first step in using a cognitive model of item writing expertise to inform the development of *design patterns* by exploring the knowledge, skills, and cognitive processes used by experienced item writers. Expert item writers who consistently write high quality figural response science items were studied using protocol analysis methods. The desired outcome of this study was to identify thought processes, knowledge, and skills that are common to these expert item writers. Once identified, these common qualities could be incorporated into *design patterns* and *task templates* supporting figural response item writing. Furthermore, we view this study as an initial step in developing a cognitive model of expertise in item writing.

This report has four sections. The first section presents a preliminary theory of item writing that serves to organize the protocol analyses and as a baseline for modeling item writers' cognition. The second section presents a study of expert item writers. The third section presents a revised model of item writing, and the final section discusses the implications for *design patterns*.

## *2.0 A Preliminary Theory of Item Writing*

Item writers are assigned an item or set of items to write. This assignment is done within the context of a number of constraints — e.g., items should be aligned to given content strands and conform to a format. Sometimes items can be written step-by-step. This kind of problem solving is described by Anderson and Lebiere (1998) and Newell and Simon (1972).

More often, item writing is best described as insight problem solving. Insight problem solving was first identified in the context of scientific discovery (von Helmholtz, 1896; Poincare, 1908/1952). Insight problem solving is characterized by an initial period of purposeful problem solving activity that may result in the rapid completion of the solution. But sometimes the item writer experiences an impasse — a state of mind in which the item writer feels that all options have been explored and he or she cannot think of what to do next. The item writer's continued concentration on the problem often causes a new idea or option to come to mind. This so-called "aha" experience is typically unanticipated by the item writer and is followed by rapid progress until the next impasse is encountered or the item is drafted.

Past theorists have described insight problem solving as an unfolding of a series of processes (Ash & Wiley, 2006; Ohlsson, 1992a; 1992b; 1996). These problem solving processes consist of three main phases: an initial representation phase, in which the solver represents the problem; an initial search through the problem space that may lead to a solution or an impasse; and, if an impasse is encountered, a post–impasse restructuring phase.

The item writer uses the writing assignment to create an initial problem representation. The initial representation and past experience activate potentially useful knowledge elements such as categories, chunks, concepts, constraints, methods, operators, procedures, rules, and schemas. These knowledge elements implicitly define a space of possible solutions or a problem space. The item writer purposefully explores this space in search of content that represents a workable solution to the assignment. The exploration of this problem solving space may lead to the successful completion of the assignment. But the space of options circumscribed by the set of active constraints might not contain any solution to the current item writing assignment.

What is a problem solving space? A problem solving space consists of three components: a set of goals (or a goal and sub–goals), states, and operators. A goal is some desired situation such as completing the item writing assignment. Sub–goals may be set that eventually lead to the goal. States define possible stages of progress in solving the problem. Operators transform a state via some mental or physical action. A problem solving space is a collection of states and operators that are available for achieving a goal. For our purposes, the problem space for item

writing will consist of a subset of likely states because the number of possible states may be infinite.

The search through the problem space is governed by three categories of constraint (Hiraki & Suzuki, 1998): object–level, relational, and goal. The object–level constraint reflects people's natural preferences of how given objects are encoded. This constrains the selection of a specific encoding of a single object among possible alternatives. The relational constraint reflects people's natural preferences regarding how given objects are related. The goal constraint involves evaluating a match between current and desired states, and gives feedback to the constraints responsible for generating the current states. The object–level and relational constraints may jointly operate to lead problem solvers to an impasse. An impasse may be resolved by the relaxation of constraints to enlarge the problem space.

When success does not follow, this initial period of purposeful problem solving activity may be followed by an impasse, a subjective feeling by the item writer that all options have been explored and the item writer cannot think of what to do next. The item writer may pause in their problem–solving behavior. The item writing impasse may be broken by restructuring the problem representation using two mechanisms for representational change (Knoblich, Ohlsson, Haider, & Detlef, 1999): the relaxation of constraints on the solution and the decomposition of perceptual chunks. The use of these two mechanisms can change the representation of the problem. A new problem representation shifts the distribution of activation over long–term memory, possibly activating dormant but relevant knowledge elements. The effect is to alter or extend the space of possibilities that are considered. If the new problem space contains a next problem solving step, the step will be completed purposefully, quickly, and unhesitatingly (Ohlsson, 1992b).

The relaxation of constraints on the solution is one mechanism for breaking the impasse in generating item content. Impasses caused by an overly constrained solution space can be broken by relaxing some constraints and thus expanding the problem solving options. Constraint relaxation is not deliberate or voluntary but is an automatic response to an impasse, though some more experienced or talented item writers may be more easily able to relax constraints. Several constraints are likely to be active during item writing. These constraints are not all relaxed at the same time. Furthermore, some constraints are more likely to be relaxed earlier than others. The constraints of narrow scope have a higher probability of being relaxed than constraints of wide scope because the resulting revisions in the problem representation are more circumscribed.

The decomposition of perceptual chunks is a second mechanism for breaking the impasse in generating item content. Experienced item writers have developed a large repertoire of problem relevant chunks. These chunks capture through experience reoccurring patterns of features or components. The item writer cannot know which chunks acquired in past experience are relevant

for the item writing assignment. However, the application of these chunks is an automatic process. If the available chunk repertoire does not parse the problem situation in a way that is helpful vis-à-vis finding the solution, an impasse might result. This type of impasse can be broken by decomposing the inappropriate chunks into their component features and so paving the way for an alternative parse of the problem situation.

## 3.0 A Study of Item Writing Expertise

Protocol analysis techniques (Ericsson & Simon, 1993) can be applied to explore how experts perform tasks such as writing items. Verbal reports of item writers' thinking while writing test items contain information on the knowledge, strategies, and reasoning employed by the writers as they develop items. Verbal reports offer an important tool in examining how item writers develop figural response items because they provide different, more direct evidence of item writers' thinking than do other methods such as observation and post–event surveys. This section presents a study in which verbal reports from expert item writers are collected and analyzed.

### 3.1 Participants

A group of three experienced science item writers was identified. These item writers had broad knowledge of science in addition to individual areas of expertise (e.g., earth science). Furthermore, they had extensive experience developing and editing assessment content, including figural response items. Two of the item writers were female and the remaining item writer was male. The first item writer had education beyond the Master's degree, five years of teaching experience, six years experience developing large–scale assessments, and three years experience writing figural response items. The second item writer had had education beyond the Master's degree, 30 years of teaching experience, four years experience developing large–scale assessments, and three years experience writing figural response items. The third item writer had a Master's degree, 17 years of teaching experience, five years experience developing large–scale assessments, and four years experience writing figural response items.

### 3.2 Materials

A storyboard scene addressing benchmarks in the History and Nature of Science content strand was developed. The History and Nature of Science strand describes science processes and skills commonly taught in elementary and secondary science courses. The History and Nature of Science strand was selected because figural response items have historically been particularly difficult for most item writers to write for this strand.

Storyboards are products of innovative test development processes that precede the development of online scenarios. Scenarios emphasize inquiry–based learning theory and hands–on science strategies and provide students opportunities to observe a process of science and the results of an investigation or event. Benchmark–aligned items are presented within the scenario context. A storyboard is a description of the narrative, images, animation, and/or video that will be developed for a scenario. A storyboard includes short narratives and descriptions of

what the student will see. Storyboards and subsequent scenarios consist of three to five scenes. Each scene provides the context for one or more benchmark–aligned innovative items. A copy of the storyboard scene is shown in Appendix A.

In addition to the scene, each item writer received a copy of the prescribed benchmarks and a figural response item template.

### *3.3 Procedure*

The participants were tested individually in one–hour sessions. Each expert was presented with one storyboard scene and two benchmarks for that scene, as shown in Appendix A. Benchmarks were presented individually, and the order of presentation was balanced so that each of the two benchmarks appeared first and second equally often across subject matter experts. Subject matter experts' verbal reports were audio recorded as they attempted to write items for each benchmark.

The experts were instructed to read and review the scene provided to them. They then were given a description of a History and Nature of Science benchmark. Finally, they were instructed to write a figural response item aligned to the scene and the assigned benchmark. After they had completed writing the first figural response item, experts were asked to write a second item for a second History and Nature of Science benchmark. They were given a description of a second benchmark and also read aloud the description. They were instructed to create a second figural response item aligned to the scene and the assigned benchmark. A copy of the instructions given to experts is shown in Appendix B.

As they developed the items, writers were asked to think aloud, verbalizing cognitive information generated during item writing. Immediately following the think–aloud verbalization for the two items, the writers were asked to answer the following four questions:

1. When you read the scene, what stands out to you as important?
2. When you read the benchmarks, what stands out to you as important?
3. As you review the first item you wrote, how do you know that you have been successful in writing an item aligned to the scene and the first assigned benchmarks?
4. As you review the second item you wrote, how do you know that you have been successful in writing an item aligned to the scene and the second assigned benchmarks?

## *3.4 Analysis*

For verbal reports to be useful data, a systematic procedure must be described for collecting the reports, extracting information from them, and using that information to test models of item writers' thinking. The analyses occurred in four steps. First, the verbal behaviors recorded during the think–alouds, and retrospective reports were transcribed for analysis. Second, the transcripts were reviewed and edited for accuracy. Third, the transcripts were segmented into individual statements. Finally, each segment was coded into broad, general categories of problem solving.

The analyses described in this paper rest on a set of assumptions about problem solving and about the verbal reporting process (Ericsson & Simon, 1993). The analyses rest on four major theoretical assumptions. First, item writers' problem solving can be viewed as a search through a problem space, accumulating knowledge about the problem situation along the way. Second, each step in the search of the problem space involves the application of an operator, selected from a relatively small set of task–relevant operators, to knowledge held in short–term memory. The application of the operator brings new knowledge into short–term memory and moves the item writer to a new point in the problem solving space. Third, the item writer's verbalizations during the think–aloud procedure correspond to some part of the information the item writer is holding in short–term memory. Finally, the item writer's information in short–term memory, and reported during the think–aloud procedure, consists primarily of knowledge required as inputs to the operators, new knowledge produced by the application of the operators, and symbols representing active goals and sub–goals. A goal may take the form of an intent to apply an operator, in which case the protocol may contain explicit evidence for the application of operators.

### *3.4.1 Segmenting*

Following their review and editing, the protocols are segmented into individual statements that may be sentences, assertions, or propositions. If oral prose were completely grammatical, each statement would be a clause or a sentence. But statements in normal speech may be phrases or even single words.

Each of the three protocols was segmented by two raters. The raters completed training before they attempted to segment the protocols. Raters were trained together. During training, raters received an introduction to text segments and viewed an example of text that had been segmented. Raters then practiced independently, segmenting a paragraph of text and discussed any differences in the segmented text.

Following training, each rater independently segmented each of the three protocols. The two raters then met and reconciled any differences in the segmenting of the protocols. The raters discussed text on which they differed until they agreed on a final segmenting.

A measure was obtained of the reliability of the segmenting process. Reliability is generally measured by having two people independently segment all of or a sample of the protocols. The reliability of the encoding process is indicated by measures of inter–coder agreement. For the analysis here, an instance of agreement was tallied if the text segment independently generated by the raters was identical to within two words. An instance of disagreement was tallied if a set of words was marked as one segment by a rater but was marked as two or more segments by the other rater. The agreement across raters was 70.34%.

The following is an example of the segmenting of a protocol of an expert attempting to create graphics for a figural response item. Note that each line is a different segment.

> Okay. All right.
> So I'm drawing five boxes.
> And up near the top is going to be another box, that's going to be
> a stimulus box
> and it'll just have to put there or have to be sized.
> I won't worry about drawing that in.
> So I'll label these A, B, C, D
> and this,
> and it's going to go zero to 100 C
> and let's see.

### *3.4.2 Coding*

During coding, statements were associated with categories. Categories may be extracted from an examination of the statements or may be constructed a priori based on an existing model or theory. The categories used in this study arose from the proposed model of item writing described in the General Theory section. This model identifies relevant information, operators, states and goals. The data from the verbal protocols were examined for the presence of statements associated with 15 different categories. Labels and a short description for these categories are provided in Table 1. These categories are defined in greater detail in Appendix C.

**Table 1. Labels and a Short Description for Each of 15 Statement Categories**

| Statement Categories | Description |
|---|---|
| Extraneous | Statements that appear irrelevant to the assignment |
| Nonconforming | Statements relevant to the assignment that do not fit any current coding categories. |
| Meta-clarification | Examples of asking for clarification about the study procedure |
| Problem definition | Creation of an initial or subsequent problem representation that includes potentially useful knowledge elements |
| Missing information | Recognition and/or searching for clarification about the assignment |
| Backtracking | Examples of retreating toward an earlier or intermediate state or even to the beginning of the problem |
| Evaluation | Evaluation of an explorative or physical operator relative to some task requirement |
| Exploratory | Examples of applying mental operators while searching for content and actions |
| Physical | Examples of bodily applying operators while searching for content and actions |
| Schema activation | Application of mental structures drawing on past experience |
| Impasse | Statements refer to a state of mind in which the item writer feels that all options have been exhausted |
| Solution | Satisfaction of some desired goals |
| Constraining | Establishment of limits on the problem solving space |
| Relaxation | Explanation of the problem solving options |
| Decomposition | Breaking-up problem-relevant chunks that represent reoccurring problem solving patterns that allow for an alternative parse of the problem situation |

In addition, some statements could not be associated with the model–based categories. These statements were categorized as ambiguous statements or nonconforming statements. Ambiguous statements were not interpretable by the raters. Nonconforming statements were meaningful but did not readily fit into any of the categories. Nonconforming statements described problem solving that did not conform to a model or theory.

As was the case for the segmenting process, a measure was obtained of the reliability of the coding process. Reliability generally is measured by having two people independently code all of or a sample of the statements and estimating the agreement across the two raters. The same raters who segmented the text also categorized the statements. Both raters categorized 33% of the statements in Protocol 2 (83 of 251 statements). The two raters selected the same coding category for 81% of the 83 statements.

### *3.5 Results*

Initially, the frequency of different categories of statements was examined. Separate analyses were completed for each protocol. For each category, the number of statements in that category was computed and the number of statements was expressed as a percent of the total statements. Differences in frequency were examined across categories and also across protocols.

The frequency and percent of different categories of statements for each protocol are shown in Table 2. Across all three protocols, a large percentage of statements were Extraneous. For relevant statements, the greatest percentage of statements was either Exploratory or Physical operators. The second greatest percentage of statements fell under the Problem Definition category. A relatively low percentage of Impasse statements was encountered.

**Table 2. The Frequency and Percent of Different Categories of Statements**

| Categories | Protocol 1 | | Protocol 2 | | Protocol 3 | |
|---|---|---|---|---|---|---|
| | Frequency | Percent | Frequency | Percent | Frequency | Percent |
| Extraneous | 19 | 13.01% | 45 | 17.93% | 11 | 7.10% |
| Nonconforming | 1 | 0.68% | 4 | 1.59% | 11 | 7.10% |
| Meta-clarification | 2 | 1.37% | 8 | 3.19% | 1 | 0.65% |
| Problem definition | 25 | 17.12% | 37 | 14.74% | 21 | 13.55% |
| Missing information | 8 | 5.48% | 5 | 1.99% | 4 | 2.58% |
| Backtracking | 0 | 0.00% | 3 | 1.20% | 1 | 0.65% |
| Evaluation | 8 | 5.48% | 19 | 7.57% | 1 | 0.65% |
| Exploratory | 26 | 17.81% | 26 | 10.36% | 54 | 34.84% |
| Physical | 11 | 7.53% | 51 | 20.32% | 21 | 13.55% |
| Schema activation | 18 | 12.33% | 9 | 3.59% | 10 | 6.45% |
| Impasse | 6 | 4.11% | 11 | 4.38% | 1 | 0.65% |
| Solution | 8 | 5.48% | 10 | 3.98% | 5 | 3.23% |
| Constraining | 13 | 8.90% | 15 | 5.98% | 14 | 9.03% |
| Relaxation | 1 | 0.68% | 5 | 1.99% | 0 | 0.00% |
| Decomposition | 0 | 0.00% | 0 | 0.00% | 0 | 0.00% |

In addition, the placement of kinds of categories was examined. The theory predicts that certain kinds of categories will occur in certain places during item writing. Problem Definition statements should occur at the beginning of item writing. The middle of the item writing activity should be dominated by Physical and Exploratory statements. The conclusion of the item writing activity should be dominated by Evaluation and Solution statements.

The placement during item writing of kinds of categories is shown in Figures 1, 2, and 3 for Protocols 1, 2 and 3, respectively. To construct these figures, the sequence of statements was divided into 5% ordered groups, e.g., the first 5% of the statements, the following 6% to 10%, the next 11% to 15%, and so on. After excluding extraneous statements, each group was examined for a dominant kind of statement where dominant was defined as more than 50% of the statements in that group. The group was labeled as “mixed” if no dominant kind of statement existed.

The placement of kinds of categories for Protocol 1 is shown in Figure 1 (page 14). The first assignment begins with a series of Problem Definition statements. Problem Definition dominates the first 15% of the protocol. Problem Definition is followed by series of Physical and Exploratory statements. These statements are interspersed with references to Schema Activation and Constraining. The second assignment is again followed with a series of Problem Definition

statements. Problem Definition is again followed by series of Physical and Exploratory statements. This protocol also included a series of Constraining statements during the first assignment. The first assignment concluded with a set of Goal Satisfaction statements.

The placement of kinds of categories for Protocol 2 and Protocol 3 is shown in Figures 2 and 3, respectively. These two protocols show a similar pattern of statements. As was the case for Protocol 1, the first assignment begins with a series of Problem Definition statements. Again, Problem Definition dominates the first 15% of the protocol. Problem Definition is followed by series of Physical and Exploratory statements. The second assignment is again followed with a series of Problem Definition statements. Problem definition is again followed by series of Physical and Exploratory statements.

**Figure 1. Placement of Kinds of Categories for Protocol 1**

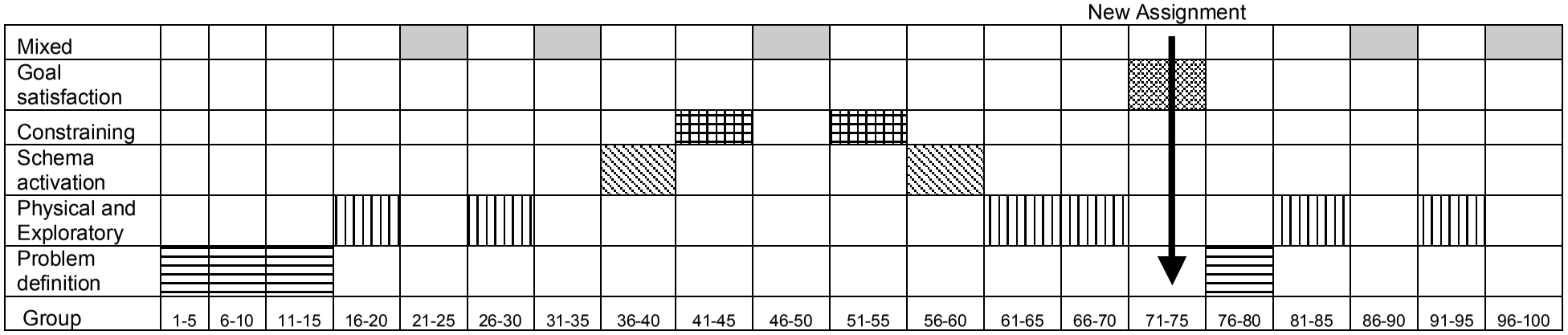



**Figure 2. Placement of Kinds of Categories for Protocol 2**

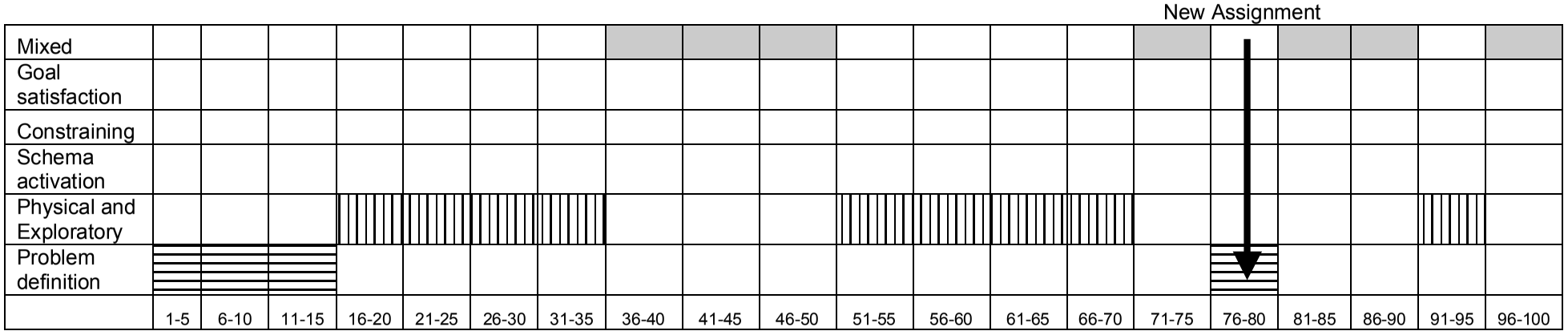
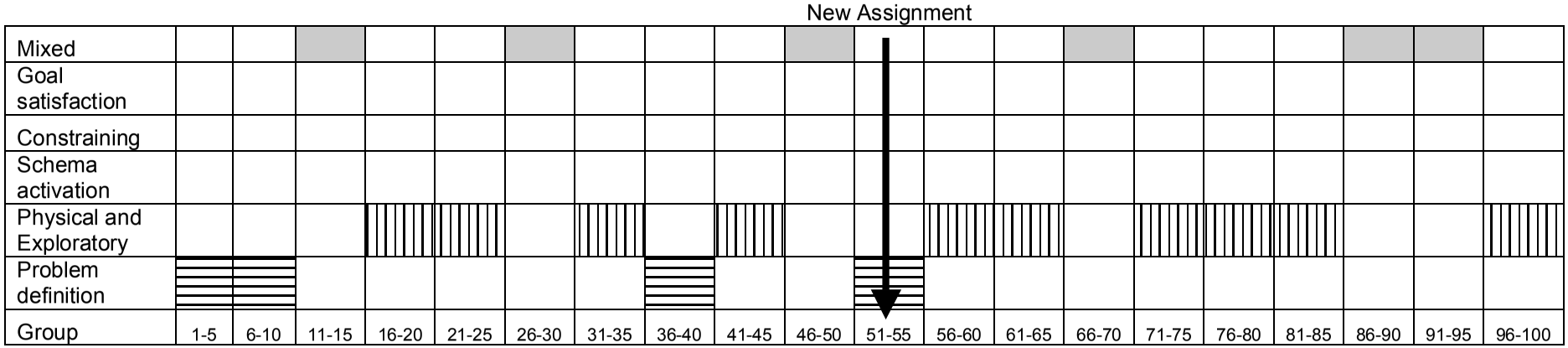


**Figure 3. Placement of Kinds of Categories for Protocol 3**



Finally, relationships among different categories were examined. The relationship was examined between the occurrence of Impasse statements and Physical and Exploratory statements. Impasse statements would be expected to occur during Physical and Exploratory statements. The relationship between the occurrence of Impasse statements and Physical and Exploratory statements is shown in Figure 4 for Protocol 1. A comparison of Figure 4 with Figure 1 shows that Impasse statements occurred during sequences of Physical and Exploratory statements or groups of Mixed statements.

In addition, the relationship was examined between the occurrence of Solution statements and Physical and Exploratory statements. Solution statements should be found interspersed throughout Physical and Exploratory statements. But Solution statements should be more frequent near the end of the applications of Physical and Exploratory statements.

The relationship between the occurrence of Solution statements and Physical and Exploratory statements is shown in Figure 4 for Protocol 1. A few Solution statements occurred during sequences of Physical and Exploratory statements. But groups with the most frequent Solution statements occurred immediately before the second assignment, at the conclusion of the first assignment, and at the conclusion of the second assignment.

The relationship between the occurrence of Impasse statements and Physical and Exploratory statements is shown in Figure 5 for Protocol 2. A comparison of Figure 5 with Figure 2 shows no clear pattern of relationships between the occurrence of Impasse statements and Physical and Exploratory statements.

The relationship between the occurrence of Solution statements and Physical and Exploratory statements is shown in Figure 5 for Protocol 2. Again, no clear relationship was apparent between Solution statements and Physical and Exploratory statements for Protocol 2.

The relationship between the occurrence of Impasse statements and Physical and Exploratory statements and between the occurrence of Solution statements and Physical and Exploratory statements is shown in Figure 6 for Protocol 3. Protocol 3 contained only a single Impasse statement and five Solution statements. The low number of these statements makes analyses difficult.

**Figure 4. Relationship between the Occurrence of Impasse and Solution Statements and Physical and Exploratory Statements for Protocol 1**

New Assignment ↓

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Physical and Exploratory | | | | | | | | | | | | | | | | | | | | |
| Solution | | | | | | | 1 | | | | | | | | 3 | | 1 | 1 | | 2 |
| Impasse | | | | | | | 1 | | 1 | | 2 | 2 | | | | | | | | |
| Group | 1-5 | 6-10 | 11-15 | 16-20 | 21-25 | 26-30 | 31-35 | 36-40 | 41-45 | 46-50 | 51-55 | 56-60 | 61-65 | 66-70 | 71-75 | 76-80 | 81-85 | 86-90 | 91-95 | 96-100 |

**Figure 5. Relationship between the Occurrence of Impasse and Solution Statements and Physical and Exploratory Statements for Protocol 2**

New Assignment ↓

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Physical and Exploratory | | | | | | | | | | | | | | | | | | | | |
| Solution | | | | 1 | | 2 | | 1 | | 1 | | | | 2 | | | | 3 | | |
| Impasse | | | | | | | | 2 | 3 | | | 1 | | | 2 | 2 | | | | 1 |
| Group | 1-5 | 6-10 | 11-15 | 16-20 | 21-25 | 26-30 | 31-35 | 36-40 | 41-45 | 46-50 | 51-55 | 56-60 | 61-65 | 66-70 | 71-75 | 76-80 | 81-85 | 86-90 | 91-95 | 96-100 |

**Figure 6. Relationship Between the Occurrence of Impasse and Solution Statements and Physical and Exploratory Statements for Protocol 3**

New Assignment ↓

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Physical and Exploratory | | | | | | | | | | | | | | | | | | | | |
| Solution | | | | | | 1 | 2 | | 1 | | 1 | | | | | | | | | |
| Impasse | | | | | | | | | | 1 | | | | | | | | | | |
| Group | 1-5 | 6-10 | 11-15 | 16-20 | 21-25 | 26-30 | 31-35 | 36-40 | 41-45 | 46-50 | 51-55 | 56-60 | 61-65 | 66-70 | 71-75 | 76-80 | 81-85 | 86-90 | 91-95 | 96-100 |

## *4.0 Revised Model of Item Writing*

Protocol data can be used to support or refute a model or theory. Verbal reports, even from a small number of subjects, can serve as evidence that at least some subjects' cognitive processes conform to a model or theory. The verbal reports serve as an important first step towards gaining evidence for generality. Similarly, evidence for the existence of some cognitive structures for some subjects can serve to challenge the universality of a theory.

Analyses from coding segments revealed that expert item writers engaged in different phases of problem solving. But the protocol data provided little evidence of different constraints item writers may encounter during problem solving (i.e., writing items to meet the assignment criteria).

Our results suggest that expert item writers engaged in three phases of problem solving. In the initial representation phase, the writers routinely created a mental model (e.g., Gentner & Stevens, 1983; Johnson-Laird, 1983) of the situation described in the scene. Problem definition statements were used to capture experts' creation of mental models. Across the three experts, problem definition accounted for between 14% and 17% of the protocol statements. Problem definition statements predominantly occurred following the item writing assignment.

The importance of this mental model was revealed by the difficulty item writers had with the term "widget" that was given in the initial storyboard scene (see Appendix A). All three expert item writers commented on the mysterious nature of the widget. Two of the three item writers created details to instantiate the widget. For example, one expert commented:

> *2 Okay,*
> *3 we're working with widgets, which are things,*
> *4 which is going to be hard to - for me to visualize…*
> *6 So in my brain I have now said that a widget is different type of metal*

In the second phase, the exploration phase, the item writer purposefully explores the problem solving space in search of content that represents a workable solution to the assignment. Exploratory and physical operator statements were used to capture experts moving toward a solution to the assignment. This phase involved a sustained sequence of processes. Across the three experts, physical and exploratory operators together accounted for between 25% and 50% of the protocol statements. The physical and exploratory operators occurred following problem definition.

Furthermore, all three expert item writers encountered at least one impasse during the series of physical and exploratory operators. For example, an item writer navigated the problem space, created a possible solution, but then realized the solution failed to satisfy all the constraints. This item writer searched the problem space again and created an alternative solution that showed no constraint violations.

In the third phase, the solution phase, the item writer successfully completes the assignment by finding a workable solution that satisfies the set of constraints. Solution statements were used to capture item writers achieving some desired situation such as completing the item writing assignment or achieving a sub-goal eventually leading to completing the assignment. Across the three experts, solution statements only accounted for between 3% and 5% of the protocol statements. The solution statements occurred following a series of physical and mental operators, toward the end of problem solving.

The conformation of all three experts to the three phases of problem solving provides evidence supporting the general theory of item writing as problem solving offered in this paper. The presence of all three kinds of statements supports the general theory of item writing. Furthermore, the sequence of statements — first problem definition, next physical and mental operators, and finally solution — conforms to the general theory.

It is noteworthy that the protocols offered little evidence with regard to impasses in item writing. Across the three experts, solution statements only accounted for between 1% and 4% of the protocol statements. The relative rarity of impasse statements may be due to the high level of expertise of the item writers in this study. A study of novice item writers or less experienced item writers might find the occurrence of many more impasse statements in their protocols.

## *5.0 Implications for Design Patterns*

With the increase in interest and use of innovative item types on large–scale assessments, test developers are experiencing pressure to quickly, efficiently, and cost-effectively produce quality innovative items for customers. The efficient development of quality innovative items is sometimes hindered by inexperienced item writers who are not familiar with the challenges and nuances of innovative item types. It is imperative, therefore, that test developers work to improve the development of innovative items.

The study of expert item writers offers the possibility of improving *design patterns* for innovative items by incorporating into *design patterns* the knowledge and skills acquired by these experts over years of hard work. This study may be one step in that process. The descriptions of the identified conceptual knowledge and skills of expert item writers could be incorporated into *design patterns* to scaffold new item writers with the knowledge and skills of expert item writers to more effectively produce quality figural response items.

An aspect of expert item writers' performance revealed by this study that could be incorporated into *design patterns* and *task templates* is the three phases of problem solving engaged in by expert item writers: the initial representation phase, the exploration phase, and the solution phase. *Design patterns* could be structured to encourage, or at least support, three stages of problem solving. Items writers could initially be encouraged to create a mental model of a situation involving a small number of scenes and evaluate the situation against a set of general constraints. Subsequently, item writers could be encouraged to explore the problem space created by the situation and instantiate the scenes. Finally, item writers could be encouraged to evaluate each instantiated scene against criteria for achieving a sub-goal or completing the assignment.

## *References*


Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Erlbaum.

Ash, I. K., & Wiley, J. (2006). The nature of restructuring in insight: an individual-differences approach. *Psychonomic Bulletin Review*, *13*(1), 66-73.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323-357). Hillsdale, NJ: Erlbaum.

Ericsson, K. A. & Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: MIT Press.

Gentner, D, & Stevens, A. L. (eds.) (1983). *Mental Models*. Hillsdale, NJ: Erlbaum

Hiraki, K. & Suzuki, H. (1998) Dynamic constraint relaxation as a theory of insight. *Bulletin of Cognitive Science*, **5**, 69-79. (In Japanese with an English abstract)

Johnson-Laird, P. N. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA: Harvard University Press. Cambridge, UK: Cambridge University Press.

Knoblich, G., Ohlsson, S., Haider, H., & Detlef, R. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *25*, 1534-1555.

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

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335-366). New York: American Council on Education and Macmillan.

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice Hall.

Ohlsson, S. (1992a). Constraint-based student modeling. *Journal of Artificial Intelligence and Education, 3,* 429-447.

Ohlsson, S. (1992b). Information-processing explanations of insight and related phenomena. In M. Keane & K. Gilhooley (Eds.), *Advances in the psychology of thinking* (pp. 1-44). London: Harvester-Wheatsheaf.

Ohlsson, S. (1996). Learning from performance errors. *Psychological Review, 103,* 241-261.

Poincare, H. (1952). *Science and method.* New York: Dover. (Original work published 1908.)

von Helmholtz, H. (1896). Vortrage und Reden [Lectures and talks, Vol. 1]. Brunswick, Germany: Friedrich Vieweg und Sohn.

Wesman, A. G. (1971). Writing the test item. In Thorndike, R. L. (Ed.), *Educational Measurement,* American Council on Education, Washington, DC.

## Appendix A

## Storyboard Scene

A student is conducting an experiment on the effect of water temperature on widget length. The student puts 1 widget 5 centimeters long in each of 5 containers of water. The water in each container is a different temperature. After 15 minutes, the student measures and records the length of each widget.

| Container | Water Temperature (in °C) | Widget Length (in cm) |
|---|---|---|
| 1 | 0 | 5 |
| 2 | 25 | 6 |
| 3 | 50 | 8 |
| 4 | 75 | 11 |
| 5 | 100 | 17 |

<table>
<tr><td colspan="2">89999_1a</td></tr>
<tr><td colspan="2">Detailed scene description: Show the data table, as given.</td></tr>
<tr><td colspan="2">Cast of Characters: Character 7</td></tr>
<tr><td>Type of Art: Still (table)</td><td>Box needed: Audio</td></tr>
<tr><td colspan="2">References for content or art:</td></tr>
<tr><td>Benchmarks</td><td>Intent/Purpose<br>(To find out if students…)</td></tr>
<tr><td>6.I.B.4</td><td>The student will present and explain data and findings from controlled experiments using multiple representations including tables, graphs, physical models, and demonstrations.</td></tr>
<tr><td>7.I.B.2, 8.I.B.3</td><td>The student will recognize that a variable is a condition that may influence the outcome of an investigation and know the importance of manipulating one variable at a time. The student will specify variables to be changed, controlled, and measured.</td></tr>
</table>

# Appendix B
# Task Instructions

## Instructions to Participants

In this study, a scene has been developed for addressing the History and Nature of Science strand. This strand was selected because figural response items have been difficult to write for this strand.

You will be asked to think aloud as you create two figural response items aligned to the scene and the assigned benchmarks. You will be presented with one scene and two benchmarks for that scene.

PRACTICE

Let's practice with a multi-column addition item. Start to think aloud as you are handed the item. Think aloud as you solve the item. Do not pause, but continue to talk.

264

+ 517

FIRST ITEM

Now that you have completed the practice task, you will be given the scene. Begin to think aloud as soon as you start to read the scene provided to you. Read aloud the scene. Next you will be given a description of a benchmark. Also read aloud the description of the benchmark. Finally, think aloud as you write a figural response item aligned to the scene and the assigned benchmark.

AFTER COMPLETING THE FIRST ITEM

Now I am asking you to write a second item. Please review the scene. Then read the second benchmark. Now write a second figural response item aligned to the scene and the assigned benchmark.

AFTER THE SECOND ITEM

Now I am going to ask you some questions. Please respond as quickly as you can, do not pause to think about the answer.

Please reread the scene. When you read the scene, what stands out to you as important?

Please reread the first benchmark. When you read the benchmark, what stands out to you as important?

Please reread the second benchmark. When you read the benchmark, what stands out to you as important?

Please review the first item you wrote. How do you know that you have been successful in writing an item aligned to the scene and the first assigned benchmarks?

Please review the second item you wrote. How do you know that you have been successful in writing an item aligned to the scene and the second assigned benchmarks?

## Appendix C
## Category Explanations

The categories used in this study arise from the proposed model of item writing described in the General Theory section. This model identifies relevant information, operators, states and goals. The data from the verbal protocols was examined for the presence of statements associated with the following categories:

1. Extraneous statements or ambiguous statements appear irrelevant to the assignment. Ambiguous statements were not interpretable by the raters.
2. Nonconforming statements are relevant to the assignment but do not fit any current coding categories. Nonconforming statements were meaningful but did not readily fit into any of the categories. Nonconforming statements described problem solving that did not conform to the current theory.
3. Meta-clarification statements are statements where the writer asks for clarification about the study procedure, not the task. An example is the following:

   *28 Do you want me to read the content (limits) out loud as I'm reading them?*
4. Problem definition statements describe the writer creating an initial or subsequent problem representation that includes potentially useful knowledge elements such as categories, chunks, concepts, constraints, methods, operators, procedures, rules and schemas. Includes reading the benchmark. An example is the following:

   *36- thinking on – back on the benchmark*
   *37 student will present and explain data and findings using representations.*
5. Missing information statements describe statements in which the writer recognizes and/or searches for clarification about the task or some aspect of the assignment.
6. Backtracking statements describes problem solving in which nothing changes in the problem representation but the writer retreats toward an earlier or intermediate state or even to the beginning of the problem.

   *99 So I have to pitch this one and backtrack and start it again on it.*

7. Evaluation statements are statements in which the writer evaluates an explorative or physical operator relative to some task requirement, constraint and/or goals

    *213The client isn't really crazy about the figural response items that are just dragging words and boxes into table, I think that - they get grumpy about that*

8. Exploratory statements are examples of applying mental operators while searching for content and actions. They may lead to the successful satisfaction of goals such as completion of the assignment or sub-goals such as defining a graphic. These statements move the writer forward in the problem space. Think of the chess master thinking forward several moves in the game without moving any pieces. These statements may be followed by an evaluative statement. An example would be:

    *68 I'm just trying to decide whether I want to have length.*

9. Physical statements are examples of bodily applying operators while searching for content and actions. They may lead to the successful satisfaction of goals such as completion of the assignment or sub-goals such as defining a graphic. These statements move the writer forward in the problem space. Think of the chess piece being bodily moved. An example is the following: REDO EXAMPLE

    *39So I guess if I were going to do this I would have an X, Y axis.*
    *40I would have –*
    *41It would be blank*
    *42and down below, that would be the stimulus*
    *44and down below would be the options that could be moved up to match the data.*

10. Schema activation statements describe drawing on past experience. Schema are structures in memory that contain slots and describe relations between slots. Schema are activated by features in the problem representation. The use of schema imposes organization onto a problem or an assignment that may or may not be helpful. Relaxation statement or decomposing statements may result to solve an impasse created by inappropriate schema. An example is the following:

*38 Well, the most obvious FR on this would be to have a graph kind of FR with three or four lines representing the data and one line correctly representing data and other lines that are generally approximated or opposites.*

11. Impasse statements refer to a state of mind in which the item writer feels that all options have been explored and the item writer cannot think of what to do next.

    *98 Okay. This is not going to work because we've got five containers and there are no intermediate values.*

12. Solution statements describe meeting some desired goals such as completing the item writing assignment. Sub-goals may be set that eventually lead to the goal.
13. Constraining statements refer to limits on the problem solving space such as the requirements that items should be aligned to given content strands and conform to a specific format. Constraining statements may be part of problem representation. Object-level constraints reflect item writers' natural preferences of how assignments are represented. Relational constraints reflect item writers' natural preferences regarding how given objects are related such as relationships between chunks, concepts, rules and methods. Finally, goal constraints involve evaluating a match between current and desired states, and give feedback to the constraints responsible for generating the current states.
14. Relaxation statements describe expanding the problem solving options by relaxing some constraints.
15. Decomposition statements refer to breaking-up problem-relevant chunks that represent reoccurring problem solving patterns that allow for an alternative parse of the problem situation.

**Sponsor**
The National Science Foundation, Grant DRL - 0733172

**Prime Grantee**
SRI International. *Center for Technology in Learning*

**Subgrantees**
University of Maryland
Pearson