

Chapter 13

Decompose, Network, Assess (DNA)

with Valerie J. Shute, Ross E. Willis, and Lisa A. Torreano

Purpose of DNA

The general purpose of the Decompose-Network-Assess (DNA) method of cognitive task analysis is to provide an easy procedure for eliciting knowledge and skill elements from experts and represent the diverse kinds of knowledge required to reason and function in any domain. The DNA method is intended to identify a hierarchically structured knowledge base of curriculum elements for instructional and training purposes (Shute, Torreano, & Willis, in press). Its primary goal is to produce the expert model for intelligent instructional systems. The DNA program is still being refined. We believe that it will be a worthwhile tool for analyzing instructional or training requirements for other, more general purposes as well.

Overview of DNA

Background of DNA

DNA is an automated cognitive tool designed to aid in knowledge elicitation and organization for instruction – particularly in relation to intelligent tutoring system (ITS) development. In addition, DNA was designed to interface with a student-modeling paradigm called SMART (Student Modeling Approach for Responsive Tutoring (Shute, 1995)). The two work in concert such that DNA extracts and organizes knowledge and skills from subject-matter experts and SMART uses the resulting knowledge structure as the basis for assessment (i.e., cognitive diagnosis) and instruction. In other words, DNA provides the blueprint for instruction, obtaining curriculum elements directly from the responses and actions of multiple subject-matter experts who answer structured queries posed by the computer (Shute et al., in press). Then the student modeling paradigm (SMART) assesses learner performance on, or comprehension of, each curriculum element by way of a series of regression equations that are based on the level of assistance the computer gives each person, per element (Shute, 1995). Thus, DNA relates to the “what” to teach, while SMART addresses the “when” and “how” to teach it.

The two specific goals of DNA are to (a) maximize the range of domains that can be analyzed, and (b) optimize the cost-benefit ratio of the process. With regard to the first goal, DNA approaches cognitive task analysis (CTA) from a perspective that focuses on the development of intelligent instructional software. Thus, the method abandons the typical restriction that the topic of analysis be a “task.” Instead, CTA is viewed as any systematic decomposition of a domain in terms of constituent knowledge and skill elements. In accord with this view, CTA may be used to analyze the knowledge structure of any domain, whether related to task performance (e.g., troubleshooting car engine problems) or not (e.g., understanding the core concepts of existentialism). To achieve this breadth of knowledge representation, DNA employs a hybrid output structure involving a mixture of semantic net and production system architectures.

The second goal of DNA is to optimize the cost-benefit ratio of doing cognitive task analysis. DNA accomplishes this goal by automating many of the time-intensive processes that are part of traditional task analysis. For example, since the SME interacts directly with the computer to delineate concepts and procedures related to his/her specific area of exper-

tise, this reduces the need for extensive transcription of protocols or coding of observational notes by the knowledge engineer. Another related way in which efficiency is improved is by decreasing the personnel resources (and hence, time and cost) required in the analysis. Traditional CTA consists of two distinct phases — elicitation of knowledge and skills, and the organization of those elements. These phases customarily occur at different points in time, and often, with different persons doing the elicitation and organization. For example, a knowledge engineer interviews or observes a subject-matter expert (SME) while a cognitive psychologist or instructional designer takes the output and arranges it into a conceptual graph or production system. With DNA, these two phases are combined into a symbiotic process in order to decrease the time and cost associated with conducting two separate analyses, both of which are massive consumers of time. In DNA, the SME identifies all curriculum elements and then arranges them into a hierarchical structure.

Description of a DNA Analysis

DNA is embodied in a series of interactive computer programs that are used first by an instructional designer, then by a subject-matter expert, and finally by a panel of experts. Information collected at each stage of the process provides a structure and database for subsequent activities.

The main modules of DNA are Decompose, Network, and Assess. However, the instructional designer initiates the domain analysis by using a Customize module. In that module, he specifies the domain (e.g., measures of central tendency), learner population (e.g., no prior statistics courses), as well as superordinate goals of the training or instructional course (e.g., know the definitions and formulas for the three measures of central tendency and be able to compute them). Additionally, the instructional designer indicates, by adjusting three “what, how, and why” gauges, the relative percentage of desired instructional emphasis or flavor of the curriculum. For instance, the instructional designer may want his experts to focus primarily on providing procedural knowledge (75%) for some training regime, with less symbolic (20%) and conceptual (5%) knowledge delineation. After obtaining all of this information from the instructional designer, the Customize module generates a brief introduction letter addressed to prospective experts and a set of floppy diskettes that contain all the necessary program files to execute DNA. The introduction letter and diskettes are forwarded to one or more experts who will use DNA to delineate the curriculum.

Decompose Module. After the expert installs DNA on his/her computer, she/he begins the Decompose module by answering a series of “What, How, and Why” questions that originate from the instructional designer in the Customize module. These questions, in general, map on to three main types of knowledge that DNA seeks to elicit: symbolic, procedural, and conceptual knowledge (for more on these knowledge types, see Shute, 1995). Decomposition can occur in a depth- or breadth-first manner, depending on the SME’s preference, at a given point in time. For instance, if the domain were “measures of central tendency,” the expert could delineate, depth-first, all elements related to the median, then go back up and do the same for the mean and later, the mode. Alternatively, she could proceed in a more global, breadth-first manner in her description of the three measures.

All questions are posed to the SME in a semi-structured interview style, and follow-on queries incorporate the expert’s responses from earlier questions. For example, if the expert identifies some procedure (X), the initial follow-on question would ask: “What is the first step you do in relation to X?” Similarly, if the expert identifies some concept (Y), one of the follow-on questions asks: “What is a typical situation involving Y?” These queries seek to obtain more information per curriculum element.

Each of the “What, How, and Why” questions has its own particular path of interrogation. Suppose an expert chose to answer one of the symbolic (“What”) questions, such

as "Define or identify a data distribution." She or he would be guided through a series of questions that aim to elicit terms and definitions related to that element. Multimedia files may be included to further embellish curricular elements. For instance, the expert could draw (with a paint program) various types of distributions to supplement the definition of distribution types. That file would then become part of the particular curricular element description. Each path is completed when the expert clicks on "Finished" and the expert is returned to the Main Question queue.

To illustrate the procedural pathway, suppose the expert chooses to answer: "What are the steps you go through when you calculate the mean?" She or he would be guided through a series of screens that allow her to delineate the steps and any conditional statements embodied within a procedure. An expert's procedure might look like the following:

- IF all frequencies in the distribution = 1
 THEN (1) sum all of the scores in the data set (i.e., ΣX)
 (2) count the total number of scores (i.e., N)
 (3) divide the summed scores by the total number of scores (i.e., $\Sigma X/N$).

Furthermore, while building a procedure, the expert is given the option to define terms that may be ambiguous to novices; thus providing additional symbolic knowledge. She or he may also develop subprocedures, group (and ungroup) co-occurring elements to disambiguate them, rearrange steps, and so on. Figure 13.1 shows the interface (i.e., the "Step Editor") that corresponds to an expert's summary of the steps underlying the computation of the mean described above.

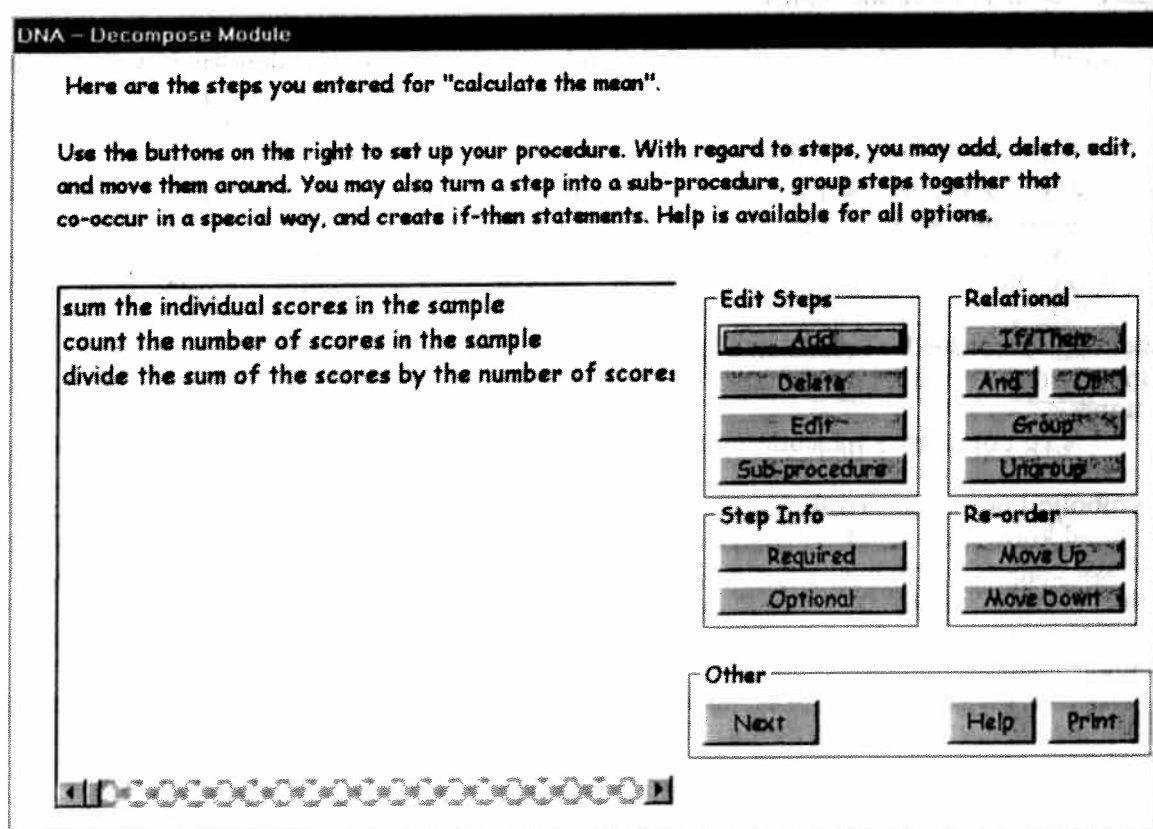


FIG. 13.1. Screen shot from DNA program.

If the expert wished to decompose conceptual aspects of the mean, she or he would be guided through a series of questions that attempt to elicit as much information about that concept as possible. Responses are typed directly into a text box that holds up to 16,000 characters. In the example domain of central tendency, the first question the expert sees is "What are the important components or issues that relate to the mean and its underlying distribution?" This question is intended to obtain an initial listing of important symbolic knowledge elements associated with the mean and various distribution types—such as normal, skewed, and bimodal. The second question in this line of inquiry is "How are these elements functionally related?" This question is designed to elicit conceptual knowledge concerning how the important components (cited in the previous response) function together. The third question is "Why is knowing about the relationship between the mean and its underlying distribution important?" This question attempts to link the current element to the overall learning goal of the instruction; again, providing the database with additional conceptual knowledge. Finally, the expert is asked to describe typical and atypical situations where knowing or understanding the relationship(s) between the mean and different underlying distributions is useful. Responses to this question supply even more conceptual knowledge.

How long do experts continue decomposing? The stopping point is indicated in the letter generated by the Customize module. That is, the instructional designer specifies the list of presumed prerequisite knowledge and skills of the intended learner population that, in turn, informs the expert as to how detailed her decomposition should be – the stopping point. Similarly, the instructional designer specifies the *ultimate learning goals* of the curriculum. These indicate the starting point for decomposition. Thus the highest- and lowest-level nodes are parameterized within the letter to the expert—providing the scope of required explication of the domain.

Network Module. The Network module loads all of the elements identified by the SME in the Decompose module and enables the expert to arrange and link graphical nodes (representing the different elements). This arrangement allows for the formation of knowledge hierarchies (similar to hierarchical task analysis, see Chapter 8), conceptual graphs (similar to conceptual graphs, see Chapter 20), or production rules (similar to cognitive simulations, see Chapter 14).

Each node contains the name of the CE and its contents as defined during the Decompose module. To simplify viewing and editing, only main-level CEs and their first-level "children" (nodes) appear upon the initial screen. "Pregnant" CEs are those that have elements embedded within them. They appear in bold font. Any pregnant element can be unpacked to reveal its components by right clicking on the node and choosing the option "unpack."

To compose a meaningful hierarchy, nodes and linkages among them differ along certain dimensions. Node shapes indicate the various knowledge types—rectangles reflect symbolic knowledge, ovals are procedural elements, and rounded rectangles denote conceptual knowledge. Links differ along four dimensions: level, type, strength, and direction of association. Some links are already in place when the SME arrives at the Network module. These come from information provided during the Decompose module (e.g., IF-THEN relationships from the "Step-Editor Window"). Other links must be drawn and labeled.

Of the links that must be made by the SME, the first relates to the *level* of relationship between two or more nodes. This establishes the inheritance hierarchy that is important both for semantic nets and procedural rules. The three level options include: parent, sibling, and child. The second kind of link relationship is *type* (e.g., is a, causes, fixed serial order). These denote the specific kind of relationship(s) between nodes. DNA's link types can relate to both semantic and procedural knowledge elements. Semantically-oriented link types allow the SME to specify the relationships among curricular elements, allowing for the conceptual structure of the domain to be specified while more proce-

durally-oriented link types allow the SME to specify the relationships among procedural steps and substeps, similar to a production-system representation. In addition to the semantic and procedural links available, there is a user-definable link that allows the SME to type in a label for a relationship not already defined. Third, links can differ in terms of the *strength* of association. There are three values for this trait: weak, moderate, and strong. This indicates the degree to which the items are related. The information on strength is accomplished by varying the width of the link line (fine, medium, and bold). Finally, the fourth link-label option is *directionality*. This refers to the flow of control or causation between curricular elements. Three options exist for this: unidirectional, bidirectional, and no direction. These relationships are established via arrowheads that are attached to the end of a line.

The use of a graphical representation should make relationships among knowledge units salient, which can also highlight missing knowledge components. This module is similar to conceptual graph analysis (see Chapter 20) except that, with DNA, experts generate the conceptual graphs instead of the instructional designers. Thus, we speculate that DNA will enable experts to recognize gaps in the knowledge and skills they provided earlier. Moreover, they have a chance to readily correct inadequacies as they can return to the Decompose module and update the curricular element record with new information.

After Smells complete the Network module, data are stored on floppy diskettes and returned to the instructional designer who reviews the curricular element record and conceptual graphs for any glaring omissions in content. If any omissions are present, the instructional designer can ask the expert to expand the inadequate curricular elements.

Assess Module. The Assess module is used to distribute the hierarchies and conceptual graphs to other experts who review and edit the database listing of curricular elements and graphs in order to validate these knowledge structures. This module is still in the design phase.

Example of DNA Output

Following are three excerpts from a CE database produced by a SME using the Decompose module. They have been only slightly edited to improve readability. During her three hours of interacting with the program, this expert explicated symbolic, procedural, and conceptual knowledge related to issues of central tendency in the domain of statistics.

Symbolic Example. *Can you define or identify the three measures of central tendency?*

CE #	Name	Description
3.001	Mode	The mode is the most frequent score in a distribution of scores.
3.002	Median	The median is the point on the scale of a distribution of scores below which 50% of the cases fall.
3.003	Mean	The mean is the average score in the distribution and is equal to the sum of the scores divided by the number of scores.

Procedural Example. *What are the steps you go through when you calculate the mean?*

CE #	Name	Description
4.001	Sum X (ΣX)	Sum the individual scores in the sample
4.002	Compute N	Count the number of scores in the sample
4.003	divide by N	Divide the sum of the scores by the number of scores (ΣX)/N

Conceptual Example. *What can you tell me about the relationship(s) between each measure of central tendency and different underlying distributions?*

CE #	Name	Description
5.001	What is the functional relationship between the <i>mean</i> and its underlying distribution?	The formula for the mean takes into account the values of all individual scores and thus is more affected by extreme scores than the median or mode. The most appropriate measure of central tendency in a particular situation also depends on the scale of measurement used. That is, the mean is used with interval and ratio data and is the preferred measure because it's the most accurate (takes into account all scores in the sample). Finally, the shape of the distribution influences the choice of a measure of central tendency. In a normally distributed sample, the mean median and mode are equal to each other. But in a skewed distribution, the mean is located closer to the tail of the distribution than the mode since extreme scores are given more weight in the formula for the mean. The median will lie somewhere between the mode and the mean in the skewed distribution. This means that the median is often a more appropriate measure of central tendency when you have a skewed distribution.

Evaluation of DNA

DNA is still under development. However, one exploratory study has recently been completed using the Decompose module (see Shute, Torreano, & Willis, 1998, for details of this evaluation). Briefly, DNA was used with three statistical experts who interacted with the Decompose module to explicate their knowledge structures related to measures of central tendency. Although experts were not given time constraints, each completed the task in less than four hours. Their output data were compared to an existing database underlying an intelligent tutor in the same domain (i.e., one of the Stat Lady modules, DS-2; Shute, Gawlick, & Lefort, 1996). The curriculum elements that were produced by all three experts were combined to determine the degree of total overlap with the Stat Lady benchmark database. Results showed that 62% of the Stat Lady curricular elements were delineated by at least one of the three experts. Thus, the agreement between the aggregate and benchmark data showed that DNA could capture a large percentage of the curricular elements present in an existing database in a reasonable amount of time.

In summary, DNA successfully achieved the rather limited goal of the pilot test. That is, it accomplished the task of eliciting relevant knowledge and skill elements from individuals, and did so as a standalone program. Further, this was achieved in hours compared to days or months with conventional elicitation procedures. These data provide preliminary information about the efficacy of DNA as a knowledge elicitation tool. That is,

given limited direction via one introductory letter of expectations for the decomposition of the domain, and minimal guidance in use of the DNA program, experts appear to be able to use the tool to explicate their knowledge structures. Moreover, the obtained data are consistent with an existing curriculum. Thus, there is suggestive evidence that DNA has potential value as an automated knowledge elicitation tool.

Advantages of DNA

There are a number of potential advantages to the DNA method of obtaining knowledge structures compared to other analysis techniques:

- Because DNA is automated, it has the potential to greatly accelerate the knowledge acquisition and organization processes that typically require exorbitant amounts of time.
- Obtaining expert knowledge structures on a domain is relatively fast; the interview, transcription, and organization processes are all automated.
- Personnel resources are reduced given that the same expert both explicates and organizes their knowledge within the same elicitation session. This contrasts with using different persons at different points in time, as is the case with traditional CTA techniques.
- DNA is theoretically grounded in an instructional framework (SMART) that uses the information from DNA-produced hierarchical knowledge structures to direct the flow of instruction.
- DNA's Decompose module utilizes what, how, and why questions that directly map onto the instructional framework of symbolic, procedural, and conceptual knowledge types embodied by SMART. These different knowledge types are associated with different instruction and assessment techniques.
- DNA's ability to obtain different knowledge types facilitates SMART's management of more customized instruction and hence expedites the development of intelligent instructional systems across a variety of domains.
- DNA is a self-contained program that conducts the interview and transcription processes. Thus, instructional designers do not have to be trained extensively in CTA methodologies to conduct an analysis (high usability).

Disadvantages of DNA

- DNA was designed to fill a particular niche—that of providing the knowledge structure (or domain expertise) for intelligent instructional systems. In contrast, the primary purpose for conducting a traditional cognitive task analysis is to delineate an expert's *performance* in relation to some task, down to a fairly small grain size (e.g., elementary cognitive processes). Given DNA's purpose of developing curriculum for intelligent instructional systems across a broad range of topics, the analysis techniques in DNA apply to domains that are based more on knowledge states than cognitive processes. Other CTA procedures may be more appropriate for defining and modeling cognitive processes underlying a particular task.
- Experts often find it difficult to verbalize much of their knowledge (Durkin, 1994). Knowledge that experts can use but cannot verbally express is often referred to as *automated knowledge* (Anderson, 1992), *tacit knowledge*, or *compiled knowledge*. In an at-

tempt to obtain automated knowledge, Durkin recommends using a CTA technique that utilizes a think-aloud protocol (i.e., ask the SME to think-aloud while performing a task). Since DNA does not currently require experts to actively participate in their domain of expertise or provide a method of capturing think-aloud data, it may not be the optimal vehicle for accessing automated knowledge.

- DNA lacks human interaction. Human interviewers can determine, in real-time, where ambiguities exist and ask experts to provide additional information as needed. Additionally, a human interviewer can give verbal and nonverbal feedback to an expert that can motivate the expert to share more information. Finally, human interviewers can, in real-time, focus on information that is directly relevant to the project and exclude information that is irrelevant. Because DNA is automated, the instructional designer must determine, after the interview, which information is deficient and which information units are relevant and irrelevant.

References

- Anderson, J. R. (1992). Automaticity and the ACT* theory. *American Journal of Psychology*, 105, 165-180.
- Durkin, J. (1994). Knowledge acquisition. In J. Durkin (Ed.), *Expert systems: Design and development*, (pp. 518-599). New York, NY: Macmillan.
- Shute, V. J. (1995). SMART: Student modeling approach for responsive tutoring. *User Modeling and User-Adapted Interaction*, 5, 1-44.
- Shute, V. J., Gawlick, L. A., & Lefort, N. K. (1996). *Stat Lady: Descriptive Statistics Module 2*. [Unpublished computer program]. Brooks Air Force Base, TX: Armstrong Laboratory.
- Shute, V. J., Torreano, L. A., & Willis, R. E. (1998). DNA: Uncorking the bottleneck in knowledge elicitation and organization. *Proceedings of ITS 98*. San Antonio, TX: Springer-Verlag (pp. 146-155).
- Shute, V. J. & Torreano, L. A., & Willis, R. E. (in press). DNA: Towards an automated knowledge elicitation and organization tool. In S. L. Lajoie (Ed.) *Cognitive tools: The next generation*. Hillsdale, NJ: Lawrence Erlbaum Associates.