36-401 Group Project Descriptions

Presented Thursday, October 25, 2001 and Tuesday, October 30, 2001

Development of Procedural Learning from Late Childhood to Adulthood

Although gross brain morphology and the basic elements of cognition are in place by school age, there is significant brain and cognitive development that continues throughout adolescence. During this time, unnecessary neuronal contacts (synapses) in the brain are eliminated and the axons become increasingly insulated through myelination, which facilitates neuronal communication. Concurrently, our ability to voluntarily control our behavior by improved working memory, response inhibition, and learning capabilities permit vast enhancement in abstract thought and cognitive planning and flexibility that define adult-level cognition. This is also the developmental period of breakdown in reasoning capabilities in severe psychiatric disorders including schizophrenia and mood disorders. In order to investigate abnormal development, however, we need to understand normal cognitive and brain maturation.

In the following study, we tested subjects throughout the span starting at 8 years of age with a procedural learning task that assesses the integrity of frontal brain regions known to be involved in basic learning. Our hypothesis is that procedural learning will continue to improve throughout childhood and adolescence reflecting the maturation of brain mechanisms known to be occurring at this time in development. Procedural learning is an implicit, relatively automatic process that reflects the cogniture of the task to speed reaction time. Latencies in unpredictable tasks are typically on the order of 200 msec; hence predictable eye movements would be faster. When individuals initiate anticipatory saccades, saccade latencies are reduced to the point where they are made either before or shortly after target appearance. This reduction in latency after learning has occurred represents a categorical shift in the nature of responses, rather than just a progressive reduction in latency, because responses become cognitively generated rather than sensorially elicited.

- Make a profile for each subject of the latency to make eye movements for each trial and determine when there was a significant reduction in reaction time.
- Are there developmental differences as to when (on what trial) "learning" occurred, i.e., there was a significant dip in reaction time?
- Overall, is there a reduction in reaction time across age?
- Is there a difference between which side the stimulus was presented? Did this difference change with age? (i.e., A left visual field preference common in adults, due to right hemisphere visuo-spatial preference, may not come on line until adolescence).

Contact person: Beatriz Luna, Assistant Professor of Psychiatry, and Director of the Laboratory of Neurocognitive Development, University of Pittsburgh Medical Center, (412) 383-1023, `lunab@msx.upmc.edu`

Classifying Readers by Speed and Accuracy

Charles Perfetti and Leslie Hart of the Learning Research and Development Center (LRDC), University of Pittsburgh, are cognitive psychologists who study reading. Reading is a fundamental skill in modern society, yet some adults can read well and others only poorly. Cognitive psychologists are interested in identifying the differences in reading processes among adults: what skills do good readers possess that poor readers lack, or vice-versa?

One of the major theories of individual differences in reading skill is known as the "lexical quality hypothesis". According to this hypothesis, reading of individual words can be broken down into three components:

- An *orthography component (OR)*, consisting of information on how the word is spelled, e.g. C–A–T;
- A *phonetic component (PH)*, consisting of information on how the word is pronounced, e.g. [kæt];
- A *semantic component (SE)*, consisting of information about the meaning of the word, e.g. whatever it is that "cat" means.

According to the lexical quality hypothesis, good readers have rich representations of words using all three components, and poor readers have poor representations in at least one component. For example, a poor reader may be able to pronounce "incarcerate" (rich PH) but have only a vague notion that the word has a negative connotation, without being able to give an exact definition (poor SE).

In this study, Perfetti and Hart administered paper and pencil tests to some 445 incoming freshmen at the University of Pittsburgh. The students were all registered for an introductory psychology course, and were required to participate in a psychology study for credit in the class. Seven tests were administered:

- Some tests were sensitive to individual components OR, PH or SE; and others were sensitive to combinations of these components.
- Test scores were recorded for both speed (either time taken to complete the entire test, or number of test questions tried in a fixed period of time) and accuracy (proportion correct of those questions tried).
- Additional measures of attention deficit disorder were also taken, since this was thought to be a possible threat to reading facility in some adults.
- In all, there are approximately 22 variables for each subject, measuring *speed* and *accuracy* at various reading skills related to OR, PH, and SE; or levels of ADD.

One test, a version of the Nelson-Denny (ND) reading comprehension test, was used to categorize readers as poor, average and good readers (e.g. if your ND score was below such-and-such, you are a poor reader, if it was above such-and-such but below such-and-such, you are an average reader, etc.). In the comprehension part of the ND test, examinees read several complete passages of text, and answer series of questions about each passage.

This project begins with two questions (though there could be many more):

- Can some combination of the other speed and accuracy variables collected in this study do a good job at predicting the good, average and poor reading skill categories that the researchers devised using the Nelson-Denny test scores? This would be useful in several ways:
  - If the other tests can be used to predict ND scores or reading skill categories, this would confirm the usefulness of the reading skill categories based on the ND scores alone;
  - By examining the components (OR, PH, SE) measured by a combination of variables that can predict ND scores or reading skill categories, we might learn which components matter "the most" in determining reading skill;
  - If the other tests are cheaper to give than the ND, then using them would provide a cheaper way to categorize reading skills of readers.
- The good, average and poor reading skill categories are primarily intended to be used in other studies. For example, in a study of what makes some words hard to distinguish from other words, we might use reading skill category as a blocking factor. However, this will only be useful if the classification consistently classifies poor readers as poor, average readers as average, and good readers as good. Can the variables collected in this study be combined in some way to produce more-consistent classification of readers than the ND test scores alone?

Contact person: Brian Junker, Department of Statistics, (412) 268-8874, `brian@stat.cmu.edu`.

Event-Related Potentials and Homophone Confusion

This is another study by Perfetti and Hart of the Learning Research and Development Center (LRDC), University of Pittsburgh, related to the "lexical quality hypothesis", a theory of individual differences in reading skill.

- An *orthography component (OR)*, consisting of information on how the word is spelled, e.g. C–A–T;
- A *phonetic component (PH)*, consisting of information on how the word is pronounced, e.g. [kæt];
- A *semantic component (SE)*, consisting of information about the meaning of the word, e.g. whatever it is that "cat" means.

According to the lexical quality hypothesis, good readers have rich representations of words using all three components, and poor readers have poor representations in at least one component. For example, a poor reader may be able to pronounce "incarcerate" (rich PH) but have only a vague notion that the word has a negative meaning, without being able to say precisely what it means (poor SE).

In this study, 48 subjects were categorized as good, average or poor readers using a version of the Nelson-Denny (ND) reading comprehension test. Then the subjects participated in a word comparison study. Each subject was presented with several pairs of words, and asked to say whether the words have related meanings (SE). The first word in each pair might be low or high frequency, and might be a homophone (sound like a differently-spelled word with a different meaning). Here is an example of each word-pair condition:

		First word		Second word
First word	Second word	Frequency	Homophone	Related meaning
Knight	Dark	Low	Yes	No
Knight	Royalty	Low	Yes	Yes
Night	Royalty	High	Yes	No
Night	Dark	High	Yes	Yes
King	Dark	Low	No	No
King	Royalty	Low	No	Yes
Evening	Royalty	High	No	No
Evening	Dark	High	No	Yes

Each subject saw 52 word pairs per condition above, for a total of 416 word pairs. In addition the amount of time between the beginning of presentation of the first word and the beginning of the presentation of the second word (stimulus onset asynchrony, or SOA) was varied both within and between subjects.

While the subjects took the test, they wore a 129 electrode cap. The cap fits over the subjects' hair and measures electroencephalograph (EEG) signals, from which event related potentials (ERP) can be derived. EEG and ERP signals measure voltage (potential) differences between each electrode and some "fixed" reference point, over time. These voltages were measured at each electrode once every 2ms over the course of approximately two seconds per word pair. Hence for each word pair and each electrode there are approximately 1000 EEG readings (there are less ERP readings since these combine EEG's over similar word pairs).

There are many questions one may ask with this data. Here are some possibilities:

- The ERP (or EEG) data from each electrode is a time series that has a "signal" (the effect of the word pair on the ERP) and "noise" (e.g. whatever other thoughts were passing through the subject's head at the moment). What is the best way to smooth out the noise, to display the "signal", i.e. differences between individuals, between word-pair conditions, between SOA's, and/or between reading levels?
- How can we summarize the information over (a) the 129 electrodes on the electrode cap; or (b) the 1000 time samples per electrode per word pair?
- Taking the first word only, are there differences in the ERP traces between word types and/or reading levels?
- Taking the second word, are there differences between word-pair conditions, between SOA's, and/or between reading levels? Are there interactions between these conditions?
- There are certain landmark "peaks" and "valley" in EEG or ERP time series: for example, a dip at 400ms may be generated when a subject is trying to resolve contradictory information, etc. Are there differences between word-pairs in terms of the heights of these peaks and valleys, or their exact timing (not exactly 400ms but maybe 390ms or 415ms, etc.)?

Contact person: Brian Junker, Department of Statistics, (412) 268-8874, `brian@stat.cmu.edu`.

National Center for Health Statistics Health Interview Survey

The National Health Interview Survey (NHIS) is the nation's primary source of general health information for the resident civilian noninstitutionalized population. NHIS is conducted by the National Center for Health Statistics (NCHS), a component of the Centers for Disease Control and Prevention, U.S. Public Health Service, Department of Health and Human Services. In accordance with specifications established by NCHS, the U.S. Bureau of the Census, under a contractual relationship, participates in the planning for NHIS and the collection of data.

Summary reports and reports on special topics for each year's data are prepared by NCHS' Division of Health Interview Statistics for publication in Series 10 of the *Vital and Health Statistics* publications series. In these reports, basic NHIS survey estimates are published annually for various population subgroups. These subgroups include those defined by age, sex, race, family income, census region, place of residence, and other domains covered on the particular NHIS statistic.

NCHS releases annual NHIS microdata files in several forms, including being made available on the internet. These files usually contain information that allows users to develop direct estimates of sampling errors that are consistent with the NHIS survey design. To maintain respondent confidentiality in NHIS, NCHS withholds variables from the NHIS's public use data files that could permit explicit or implicit identification of survey respondents. One of the major risks for inadvertent respondent identification is the inclusion of identifiers on survey files that place respondents in small geographic areas. Thus, in these data releases, variables identifying specific geographic areas smaller than one of the four census regions usually are withheld to protect respondent confidentiality. However, NCHS has released information that allows identification of the largest metropolitan areas.

Note - Summary obtained from NCHS NHIS manual and http://www.cdc.gov/nchs/nhis.htm

Contact person: Steve Fienberg, Maurice Falk University Professor of Statistics and Social Science, (412) 268-2723, `fienberg@stat.cmu.edu`

Longitudinal Long-term Care and Disability analysis

Changes in the prevalence of chronic disability in the United States elderly population are important for several reasons. For example, declines in chronic disability over time can have major implications for fiscal stability of Medicare and Social Security programs. Documented chronic disability declines among the U.S. elderly population between 1982 and 1994 led to projections about the positive consequences of continued and even accelerating declines. More generally, many models of health changes in developed societies have suggested that industrialized, economically advanced countries would present social and public health problems that increase chronic disease and mortality risks. Secondly, chronic disability, especially for elderly populations, is a sensitive measure of age-related changes in the health and biological fitness of individuals.

The arguments for using disability measure to make these assessments are substantive and methodological. Substantively, the dimensions of chronic disability are not only a product of chronic morbidity but are, increasingly at advanced ages, primary risk factors for diseases such as stroke, coronary heart disease, peripheral vascular disease, diabetes, and cancer. Methodologically, longitudinal measurements of chronic disability are important because they are easier to assess in nationally representative population surveys than many physiological risk factors or biomarkers of aging, which require blood drawing or other physically invasive (and expensive) tests and procedures. The information content of longitudinal population monitoring can be greatly enhanced by linking an individual's survey data to his/her continuous records of Medicare Part A and B service use. They also provide a temporal matrix for integrating data from specialized health studies– an integration that increases the value of that specialized health information by helping to assess, and possibly enhance, its population and temporal representativeness. Note that, though chronic disability time series are important to monitor age-specific changes in population health and biological fitness, measurement of chronic disability, especially in elderly populations, is complicated by the different types and degrees of disability manifest.

Note - Summary obtained from the following references:

Kenneth G. Manton and XiLiang Gu. "Changes in the prevalence of chronic disability in the United States black and nonblack population above age 65 from 1982 to 1999" PNAS 98: 6354-6359.

Kenneth G. Manton, Larry Corder, and Eric Stallard. "Chronic disability trends in elderly United States populations: 1982-1994" PNAS 94: 2593-2598.

Contact person: Steve Fienberg, Maurice Falk University Professor of Statistics and Social Science, (412) 268-2723, `fienberg@stat.cmu.edu`

Disease analysis – spatial and/or temporal model

Data is available from the Centers for Disease Control and Prevention's National Notifiable Diseases Surveillance System (NNDSS) on a variety of diseases at the state level. Weekly data is provided from 1980-1994 per state on approximately 50 notifiable diseases, including hepatitis A, hepatitis B, malaria, measles, mumps, rubella, gonorrhea, and syphilis. As a result, there exists a variety of possible projects with respect to these various diseases, such as:

- Can we describe any trends regarding the prescence of salmonella in Massachusetts? Do trends exist with respect to time?
- Creating dynamic maps (i.e., making a movie) to study the trend(s) of a certain disease(s)

In essence, the interested group can produce dynamic maps describing the spatial trend over time, or determine appropriate spatial and/or temporal models to describe the prescence of a particular disease. For example, the data associated with mumps is interesting given the case sample size and history with respect to the disease. Questions or projects of interest include the following:

- Create a temporal model to describe the spread/decline of mumps over time (in a particular state or across the country).
- Create a dynamic map (i.e., make a movie) to study the spread/decline of mumps from a spatial perspective.
- Do relationships exist between mumps and any other diseases, such as measles or rubella? Is there any correlation between them?

Contact person: Bill Eddy, Professor, Department of Statistics, (412) 268-2725, `bill@stat.cmu.edu`

Image analysis of microarray data

The Human Genome Project recently publicized that they have successfully sequenced most of the human genome. As a result, there is a vast amount of statistical data available in this new and rising field. In this project, we have been presented with five microarray experiments. These are large scale microarray experiments testing for genetic patterns among schizophrenic individuals, i.e. can we locate a genetic makeup that defines schizophrenia. There are approximately 10,000 spots per array, where the tissue analyzed on each microarray consists of a control sample and a schizophrenic sample. The control and schizophrenic samples, coming from the prefontal cortex region of the brain, have been matched on as many variables as possible.

For each spot on the microarray, we have the associated location with respect to the row and column number for each spot. Further for each spot we have a plate identification. That is, each microarray chip is created from several smaller plates, each with 96 spots representing the dye solution of cDNA. For these experiments, each microarray is created by transporting the solution from 106 plates(each plate containing 96 wells). A machine, called an arrayer, has four pens that lift the spots from the plates and place them onto the microarray chip.

Although, the initial project was to examine the data to search for relationships among schizophrenic patients, this particular project delves deeper to analyze the data itself. We want to study whether the current procedures that are quickly gaining acceptance among genetic scientists are, in fact, legitimate. Therefore, we use the data that is supplied to "recreate" a spatial representation of the microarray chip and test for any location effects (we are mainly interested in the signal effects):

- Does the location of the spot on the microarray slide (chip row and/or chip column) affect the data outcome?
- Does the location of the spot with respect to the plate number have any affect on the outcome?
- Does the location within the plate (i.e., plate row and/or plate column) affect our data?
- Does it matter which pen transports the solution from the plate to the microarray chip?
- Does the order by which the pens transport the four solutions affect the outcome data? In other words, are the pens sufficiently cleaned between transport procedures?

Unfortunately for these experiments, the same dye was used throughout on schizophrenics while the other dye was used for all of the controls. Thus, the dye effect and the sample effect are confounded. That is, we cannot differentiate between a sample effect and a dye effect in gene expression.

Contact person: Kim Sellers, Visiting Assistant Professor, Department of Statistics, (412) 268-3967, `ksellers@stat.cmu.edu`

Motivations for Caregiving

The purpose of this research project is to examine motivations for providing caregiving, and motivations for not providing caregiving, in romantic relationships. Couples who have been involved in a stable romantic relationship were recruited to participate in this study. Both members of each couple were asked to participate in a 3-part investigation. In the first phase of the study (from which this dataset emerged), each member of the couple was asked to complete a variety of questionnaires. The most relevant questionnaires for the proposed research questions are described below. Couple members completed their questionnaires in separate rooms in order to ensure complete privacy.

A Motivations for Caregiving and a Motivations for Not Caregiving questionnaire was designed specifically for use in this study. The Motivations for Caregiving questionnaire asks participants to rate the degree to which they generally help their partners for a variety of reasons (e.g., I want my partner to be happy; I want my partner to need me and depend on me). The Motivations for Not Caregiving questionnaire asks participants to rate the degree to which they generally do NOT help their partners for a variety of reasons (e.g., I don't have the time; my partner doesn't appreciate my helping efforts).

Couple members also completed a variety of caregiving measures to assess the degree to which they tend to be responsive, compulsive (too much), or controlling (too bossy) caregivers in their relationship. Couple members also completed some indices of overall relationship functioning, such as measures of relationship satisfaction and conflict. Finally, some important personality measures were included. The measures of attachment style are the most important personality variables for the research questions proposed below.

- Based on factor analyses of the Motivations for Caregiving questionnaire, can empirically valid subscales be constructed for use in data analyses (e.g., obligation motives for helping, reward motives for helping)?
- Based on factor analyses of the Motivations for Not Caregiving questionnaire, can empirically valid subscales be constructed for use in data analyses (e.g., partner too dependent motives for not helping, time/resource constraint motives for not helping)?
- Do the motivation for caregiving subscales computed above predict the quality of caregiving provided within the relationship?
- Do the motivation for NOT caregiving subscales computed above predict the quality of caregiving provided within the relationship?
- Does the couple member's attachment style predict his/her motivations for caring and for not caring for the partner?
- Do the motivation for caregiving and motivation for not caregiving subscales predict overall indices of relationship functioning (relationship satisfaction and conflict)?
- Does the quality of caregiving provided within the relationship (responsive, compulsive, controlling caregiving) mediate the link between motivations for caregiving (and for not caregiving) and relationship functioning (satisfaction, conflict)?

  Motivations for Caregiving → Caregiving Quality → Relationship Functioning

- Do motivations for caregiving mediate the link between caregiver's attachment style and caregiving quality?

  Attachment Style $\rightarrow$ Motivations $\rightarrow$ Caregiving Quality

Contact person: Brooke Feeney, Assistant Professor, Department of Psychology, (412) 268-6203, `bfeeney@andrew.cmu.edu`

Coping

The purpose of this research project is to examine the degree to which an individual's own coping style predicts the way in which he/she provides care to significant others. Both members of romantic couples (who were involved in serious dating relationships) participated in the study. Couple members first completed some background questionnaires assessing their attachment style, coping style, and general reports of relationship functioning (satisfaction, conflict).

Later in the laboratory session, one member of each couple was randomly assigned to take the role of a "support-recipient" by thinking of a current, ongoing stressor in his/her life that does not involve conflict with his/her partner, then discussing this problem with the partner. The other member of the couple was implicitly assigned to take the role of a "caregiver." Couple members were videotaped for 10 minutes as they discussed this problem. The interactions were later coded for support-seeking and caregiving content by objective raters. After the interaction, both members of the couple rated their perceptions of the interaction (e.g., how much they enjoyed it, how supportive the caregiver was).

- Does the caregiver's scores on the various coping scales predict the quality of caregiving that he/she provides to his/her partner?
- Does the caregiver's attachment characteristics predict the caregiver's scores on the various coping scales?
- Does the caregiver's coping scores mediate the link between the caregiver's attachment characteristics and the quality of caregiving he/she provides during the interaction?
  Attachment Style $\rightarrow$ Coping $\rightarrow$ Caregiving
- Does the caregiver's scores on the various coping scales predict the quality of the support-recipient's seeking of support?
- Does the caregiver's coping style predict the recipient's perceptions of the caregiver's behavior during the interaction?
- Does the caregiver's coping style predict his/her own perceptions of his/her own caregiving behavior during the interaction.
- Does the caregiver's coping style predict his/her attributions regarding the partner's problem?
- Are there gender differences in any of the above processes?

Contact person: Brooke Feeney, Assistant Professor, Department of Psychology, (412) 268-6203, `bfeeney@andrew.cmu.edu`

Merkle Direct Marketing, Inc. Consultation projects

Merkle Direct Marketing, Inc. provides database-marketing services to companies and organizations that thrive through direct marketing. In particular, the Analytical and Consulting Services group services these companies by extrapolating information from the company database to analyze customers who are responders or non-responders to their marketing tactics. The division runs "profiles" to compare the overall average response rate to the average response rate within different categories (e.g., sex, age, income).

After analyzing the results of the profile, they use various statistical packages to create and test a model to inform the company of the common traits among the company's clientele within responders and non-responders. Their model is created using 75% of the sample population to gain some understanding about the responders. After the model is created, they test the model on the remaining 25% of the sample population to check the accuracy of the model.

Note that these models are usually logistic or linear regression models designed to test the significance of any trends relating to the response rate and rank the results from highest to lowest to inform the interested party of key characteristics (and therefore key persons) to target in order to increase the company's response rates.

Below are two possible projects for interested groups:

Fundraising Model

- Client Overview: Major fundraiser whose primary goal is to educate people on nutritional issues. They have a major monthly newsletter that people can subscribe to.
- Objective: To build a model that will identify which subscribers of the newsletter will also donate to this fundraiser. The model will be applied to the entire subsubscriber base and only the top 25% mailed a donation request. In other words, the client's goal is to convert the magazine subscription universe to donators to their cause.
- Dataset: Sample of subscribers who were mailed last year and those who donated due to the mailing. There will be 100+ potential predictors including previous subscription information and demographic data. The analysis (response) variable is called "response" and is coded 1,0: 1 means that the magazine subscriber converted to a donor in the past year, and 0 means that the magazine subsriber did not convert to a donor in the past year. All of the internal transaction data is up to 2000-09 (YYYY-MM). The prediction (response) period is from 2000-10 to 2001-09.

Wealth Model

- Objective: To build a model that predicts a wealth score for each household in the USA. The model will then be applied to all households in the US.
- Dataset: We will send a sample of the US population with 50+ different predictor variables such as income, home value, age, etc. No names and addresses will be sent.

Contact persons:

Tim Berry, Vice President of Analytical Services, Merkle Direct Marketing, Inc., 301-429-4350, `tberry@merklenet.com`

Darryl Putnam, Senior Statistical Analyst of Analytical Services, Merkle Direct Marketing, Inc., 301-429-4333, `dputnam@merklenet.com`