INL/CON-07-12774
PREPRINT

A Review of Expertise and Judgment Processes for Risk Estimation

Proceedings of the European Safety and Reliability Conference (ESREL 2007)

R. L. Boring

June 2007

The INL is a
U.S. Department of Energy
National Laboratory
operated by
Battelle Energy Alliance

A Review of Expertise and Judgment Processes for Risk Estimation

R.L. Boring
Idaho National Laboratory, Idaho Falls, Idaho, USA
and
OECD Halden Reactor Project, Halden, Norway

ABSTRACT: A major challenge of risk and reliability analysis for human errors or hardware failures is the need to enlist expert opinion in areas for which adequate operational data are not available. Experts enlisted in this capacity provide probabilistic estimates of reliability, typically comprised of a measure of central tendency and uncertainty bounds. While formal guidelines for expert elicitation are readily available, they largely fail to provide a theoretical basis for expertise and judgment. This paper reviews expertise and judgment in the context of risk analysis; overviews judgment biases, the role of training, and multivariate judgments; and provides guidance on the appropriate use of atomistic and holistic judgment processes.

1 INTRODUCTION

Expert estimation or elicitation involves polling subject matter experts to produce a probability of human error or hardware failure. The analyst who orchestrates the expert elicitation incorporates expert estimates into an overall risk model. Expert elicitation has proven especially useful within safety-critical industries to ensure compliance within regulated operating parameters. When operational data for hardware or performance data for human operators are not available, expert elicitation provides quantitative measures suitable for inclusion in probabilistic safety assessment (PSA) and human reliability analysis (HRA) models. These models ensure optimal safety of systems by identifying and minimizing risks.

An expert elicitation for the purpose of supporting risk assessment involves two components: (i) a subject matter expert and (ii) a judgment about the likelihood of event occurrence. It is common to focus primarily on one of these areas; however, it is absolutely necessary to consider both. The success of expert elicitation hinges on the well-orchestrated interplay of the right subject matter expert using the right information (or the information available) in conjunction with the correct method to judge event likelihoods. This paper reviews expertise and judgment, offering new insights into how to use these components to improve the quality of expert elicitation in risk assessment.

2 CHARACTERISTICS OF EXPERTISE

Expertise involves experience and knowledge that can be applied to a particular domain. Simon and Chase (1973) suggest that for most domains it takes a minimum of ten years of experience to achieve expertise. These ten years are not prescriptive for all expertise or all individuals. New expertise may transfer directly from existing expertise and proficiency, effectively shortening the amount of time and experience that is necessary to yield expertise. It should also be noted that while having ten years or equivalent of experience may be requisite for expertise, ten years of experience by no means guarantees expertise. A large component of experiential expertise involves having had sufficient long-term exposure as well as active engagement in the domain to make logical inferences for novel situations within that domain. In the case of risk analysis, it is easy to see that a combination of experience in operations, regulation, model development, and hands-on PSA and HRA are desirable characteristics that contribute to expertise.

Expert knowledge is more coherent and more structured than novice knowledge (Wilson, 1994), with the degree of knowledge coherence and structure increasing through experience. Knowledge is a natural byproduct of experience. But, expert knowledge is only achieved through ongoing activity in the topic (Ericsson, Krampe, and Tesch-Römer, 1993). Expert knowledge requires ongoing practice in the topic of mastery coupled with a deliberate effort to increase knowledge. Simple passive exposure and experience without engagement, motiva-

tion, and effort do not craft expertise nor deep knowledge and understanding of the domain.

Anderson (1995; see also Rasmussen, 1985) suggests that there are three stages in the acquisition of skills and expertise. During the first stage—the *cognitive stage*—an individual acquires knowledge in the form of facts. During this phase, the individual has so-called "textbook learning" but has limited application of the principles he or she has learned. In the second stage of skill and expertise acquisition—the *associative stage*—the individual begins to apply and use the knowledge. The individual is an "apprentice" in synthesizing factual knowledge with application. In the third and final stage—the *autonomous stage*—the individual becomes an expert. Knowledge becomes proceduralized and automatic rather than studied or deliberate. At this stage, the expert individual is able to apply knowledge quickly and effortlessly to his or her domain of expertise.

Weiss and Shanteau (2003) argue that evaluative skill is the defining characteristic of expertise. Evaluative skill combines with a particular elicitation domain to yield topical expertise. Weiss and Shanteau present the following examples of topical expertise:

1. *evaluation + qualititative or quantititative expression = expert judgment*
2. *evaluation + projection = expert prediction*
3. *evaluation + communication = expert instruction*
4. *evaluation + execution = expert performance*

Expert elicitation within risk analysis falls into the first two categories. Expert judgment is made regarding the circumstances of an off-normal event. The analyst's evaluation of the circumstances is combined with quantification of the event likelihood to yield estimated failure rates or human error probabilities. In some cases the risk analyst must project event likelihoods given minimal actual operating data. Expertise resides in the analyst's acumen at translating domain knowledge of operations into a judgment or prediction of event occurrence. Within formal PSA, expert estimation is often accomplished through the aggregation of estimations from multiple experts or analysts. The present discussion considers expertise and judgment primarily from the perspective of a single expert.

3 JUDGMENT

For the purpose of the present discussion, judgments and predictions about event likelihoods are treated as identical processes and simply called judgments. Judgments are the actual assignment of a rank, quantity, certainty bounds, or probability to an event. There are a number of general frameworks for gauging event likelihood and for translating that likelihood into a number suitable for analysis. These methods are elaborated below.

3.1 Atomistic and Holistic Judgment

3.1.1 *Introduction*

There are two common types of quantifiable judgment processes—*atomistic* and *holistic*. Atomistic judgments involve breaking a judgment area into constituent subcomponents. Independent judgments are made about each subcomponent, and later aggregated into a summary judgment. In HRA, for example, the atomistic model of judgment is the method employed in estimating human error based on performance shaping factors. For example, in the SPAR-H method (Gertman et al., 2005), the human error probability is the sum of the influence of eight performance shaping factors on the default or nominal error rate.

In contrast, in holistic judgments, a judgment about the overall event likelihood is made. Holistic judgment eschews individual contributing factors (like performance shaping factors) but, instead, views the event and circumstances as irreducible. In risk analysis, the holistic model of judgment is the classic method of expert elicitation. In the holistic scaling methods advocated in NUREG/CR-2743 (Seaver and Stillwell, 1983) and NUREG/CR-3688 (Comer, Seaver, Stillwell, and Gaddy, 1984), analysts make probability judgments about the likelihood of the event, but the analysts do not explicitly quantify the subfactors that contribute to the overall error probability. In a sense, holistic proponents have argued that the sum is not the product of the parts but rather is the simultaneous interaction of all parts related to safe operations. Holistically, this interaction is considered irreducible.

3.1.2 *Historical Basis of Atomism and Holism*

It is important to consider the theoretical basis of the distinction between atomism and holism. The notion that there are two types of mental processes—one holistic and another atomistic—has been the source of considerable research and discussion in the psychological literature, both prior to and concurrent with their emergence in expert elicitation for event likelihood in off-normal operating events. Underlying this is the issue regarding humankind's ability to quantify—to associate probabilities with either atomistic or holistic thought.

Early in the history of experimental psychology, there was a school of psychology centered on the idea of the deconstructability of mental processes. Edward B. Titchener, chief proponent of the structuralist movement in psychology, suggested that

consciousness should be investigated in terms of reduced component structures or processes, rather than as a singular, irreducible process (Titchener, 1911). Titchener advocated the deconstruction of all human mental processes through use of systematic introspection. Although introspection ultimately proved a better resource for qualitative than quantitative analysis of thought, the importance of structuralism has endured into contemporary psychology.

It was the Gestalt psychologists, most notably Max Wertheimer, Wolfgang Köhler, and Kurt Koffka, who developed the strongest argument against structuralist analysis of thought (Köhler, 1947). Through a series of now famous perceptual illustrations, the Gestalt psychologists demonstrated how the mind operates according to grouping principles. No matter how hard a person might try to separate the components of visual objects that he or she sees, these objects are unfailingly perceived in groups or Gestalts. The key to the Gestalt argument is that this grouping process occurs automatically, as a subconscious process. In modern parlance, the process of grouping is said to be cognitively impenetrable (Pylyshyn, 1984), meaning the perceiver does not have conscious access to the mental processes that cause the perception of a single object. Moreover, because the process is cognitively impenetrable, the perceiver is not able to control the grouping processes that are at work. Normal human object perception can never disconnect those components that group together to create a single object.

At the heart of the structuralist/Gestaltist debate is atomistic and holistic theory. Titchener and other structuralists espoused an atomistic view of the mind, which is to say that a mental process is the sum of its constituent parts. Conversely, the Gestalt psychologists argued for a holistic view of mind in which a mental process is viewed as more than the sum of its parts. At issue is also whether or not thoughts are cognitively penetrable. An atomistic model of the mind holds that mental processes are comprised of discrete mental steps, which may be accessed on a conscious level. A holistic model of mind holds that mental processes are comprised of automatic processes that are not consciously accessible.

3.1.3 *Atomistic and Holistic Risk Analysis*

Atomistic and holistic expert elicitation methods are loosely based on atomistic and holistic theories of human cognition. In Gestaltist terms, the event that is being judged by the analyst cannot be decomposed into and evaluated as constituent properties. A holistic theorist would argue that an expert elicitation represents an indivisible process and that an event likelihood estimation cannot be judged merely as the sum or product of the individual performance shaping factors or degraded hardware reliability factors. To evaluate the probability of failure, the holistic analyst should look at no smaller object than the whole event itself.

In marked contrast, these elemental components would be the only basis of evaluation by an atomistic theorist. An atomistic theorist would attempt to decompose the event into its most elemental units, such as performance shaping factors. Each of these components would be evaluated individually, and the composite event likelihood would be the sum or product of the values awarded for each factor. Atomistic expert elicitation is, in fact, characterized by a formal set of procedures for merging ratings from individual contributing factors into a composite event likelihood.

It is important to note that each method of expert elicitation has documented shortcomings. Atomistic elicitation, for example, is known to fail due to insufficient information for completing the atomistic rubric or due to clerical or procedural errors in completing the atomistic forms or worksheets (Hammond, Hamm, Grassia, and Pearson, 1987). Moreover, there is serious difficulty in designing a valid and comprehensive atomistic rubric. A poorly designed atomistic rubric will hamper efforts to arrive at a meaningful representation and quantification of the problem space. Designing a solid atomistic scoring rubric is exacting and time consuming. For example, the SPAR-H worksheets (Gertman et al., 2005) represent ten years and three full iterations in terms of development history. This development time was necessary to produce a comprehensive list of performance shaping factors, map the relationship between these performance shaping factors and human error probabilities, and make adjustments based on analyst feedback. Other atomistic approaches to expert elicitation in the safety-critical arena feature comparably long development histories.

Holistic elicitation features a similar array of shortcomings. One primary concern is the level of expertise required to perform holistic elicitation. Because holistic elicitation allows the analyst to evaluate according to their own criteria and impressions, this method is not well suited for novice analysts. The use of novices in holistic elicitation results in very inconsistent ratings, as novices may use ad hoc or inoperative judgment processes (Madigan and Brosamer, 1991). A related concern with holistic elicitation is that it commonly enlists selective information about the object of investigation. Ettenson, Shanteau, and Krogstad (1987) show that expert analysts tend to focus on selective information about a problem space, to the detriment of other information. For example, one risk analyst may focus on how the failure event will fit within the plant model construction (i.e., what configurations have been modeled or what components may have been incorporated into a supercomponent, etc.), while another analyst may give more weight to his or her impres-

sion of plant conditions as opposed to model representations of those conditions. While this selectivity allows analysts to focus on the most crucial information, it may also hinder the analyst from considering other contributing factors. The holistic method does not explicitly require the analyst to make a broad sweep of the problem space. This means that the analyst, especially the novice analyst, may omit important information when considering the likelihood of an event. The open-ended evaluation criteria of holistic elicitation are a significant contributor to the holistic method's generally low inter-rater reliability as reported in Boring (2003a).

3.2 Judgment and Scaling

3.2.1 *Error Estimation Process*

Both atomistic and holistic approaches to expert elicitation involve determination of a failure rates through estimation. In atomistic approaches, a baseline is incremented or decremented using hardware degradation considerations or human performance shaping factors. In holistic approaches, the failure rate is estimated based on the overall event context. Failure estimation entails comparison either to a standard or to other tasks or errors that the analysts review. There is a qualified body of research in the psychological literature that deals with ensuring a stable, repeatable process for comparison. This process is known as psychological scaling.

Stillwell, Seaver, and Schwartz (1982) performed research for the US Nuclear Regulatory Commission in which they compared five common methods of psychological scaling. These are:

1. *Scaling by paired comparison*. In this method, all possible pairs of events are presented to the analyst. The analyst then judges for each event pair which event is more likely to occur. This method provides the frequency that a given event is more likely to occur than another event.
2. *Scaling by ranking.* In this method, the analyst assigns a rank order or event likelihood to each event, resulting in a rank ordering all event likelihoods.
3. *Scaling by sorting.* This method, also known as card sorting, entails sorting events into piles based on their similarity. In terms of failure estimation, the piles can be sorted to represent different categories of event likelihood.
4. *Scaling by rating.* This method, also known as categorical judgment or Thurstonian scaling, entails assigning an event to a category label or number that reflect the event likelihood. This method has the advantage that an event may be analyzed in isolation, without the need for a comparative listing of events.
5. *Scaling by fractionation*. This method is also known as magnitude estimation or direct estimation. In this method, the analyst assigns a number or similar descriptor to the event to indicate the event likelihood. Like scaling by rating, this method allows the analyst to look at a single event without the need for comparing other events. This method features a further advantage in that the resulting scale is a continuous scale instead of a categorical scale. The categorical scale produced by scaling by rating does not guarantee equal quantitative intervals between categories. The distinction between "medium" and "high" is not quantifiable to the same degree that the distinction between 3.5 and 4.7 is. Categorical scales are known as nominal or ordinal scales, while continuous scales are known as interval or ratio scales (Stevens, 1975).

Stillwell et al. (1982) suggest that scaling by fractionation is the most powerful method available for expert elicitation, although it is susceptible to scaling biases. By bias is meant the fact that people do not treat a scale in a consistent, linear fashion.

3.2.2 *The Role of Bias*

Poulton (1989) provides a comprehensive list of biases in psychological scaling. These include:

1. *Contraction biases.* There is, for example, a series of contraction biases, in which people do not use the full range of the scale available to them. A common form of this occurs when people scale most magnitudes too closely to a central scale value.
2. *Inappropriate units of measurement.* Other biases occur when people use inappropriate units of magnitude. Such would be the case, for example, when judging temperature on a novel scale despite high familiarity with the Fahrenheit or Celsius scale. The familiar Fahrenheit or Celsius scale could hamper the ability of a person to scale using a different temperature scale, and the person would have a proclivity for inadvertently using the familiar scale.
3. *Logarithmic response biases.* When people do not use a scale consistently across the scale range, they often exhibit a logarithmic response bias. This may happen when a person does not know how to map a magnitude to a scale in a particular range. Familiarity with one part of the scale range breeds finer granularity for that part of the scale. For example, if a person scales a stimulus in one-step increments in a low range and then uses ten-step increments in a high range, this would result in a logarithmic shaped curve.

4. *Range equalizing biases.* In some cases, people scale using the entire range of a scale, even when the stimulus does not warrant a full range of responses. In such cases, people exhibit range equalizing biases. If, for example, a subset of stimuli is presented at around the midpoint of a scale yet the person provides responses spanning the full range, then that person is equalizing the range of his or her scale response.
5. *Centering biases.* When people use all scale responses above and below the midpoint of a scale equally often, there are centering biases. For example, when a person is presented with a range of loud stimuli, that person might assign loudness values localized around the midpoint of the scale. When presented with a range of quiet stimuli, the person might assign roughly the same range of responses. Although the stimuli are clearly different, the response bias of the individual tends to center the scale responses, thereby minimizing the scale differences between the two stimuli. Helson's adaptation-level theory (1964) accounts for this phenomenon, in which people adapt their perceptual response according to the intensity of the stimuli with which they are presented.

Although Poulton's scaling biases are most clearly illustrated through perceptual scaling examples, scaling biases are endemic to expert elicitation as well. These biases most commonly manifest themselves in the form of a failure to use the full range of the probabilistic scale (i.e., a contraction bias) or a tendency inappropriately to use the extreme probabilities (i.e., a range equalizing bias). Thus, it is imperative to provide training and or instruction to experts in order to help limit bias influences.

Without guidance, analysts may fail to assign the full range of applicable probabilities and instead tend to focus on extremes, i.e., assigning events as having an unrealistically low or high probability of occurrence. Alternately, analysts may assign neutral, midrange probabilities to events that would more accurately reflect a lower or higher probability.

With these many biases coming into play, analysts need to exercise caution before treating their scaling measures with the kind of confidence that a physical scientist might exercise. Too often, analysts correct for biased scaling results without remedying the cause of the biases. Mathematical corrections are applied to the estimations and groups of analysts are enlisted for a study to compensate for scaling biases (Meyer and Booker, 1990). While there is definite value in post hoc scaling corrections, there is also a need to prevent scaling biases.

3.2.3 *The Role of Training*

The key to preventing scaling biases is training. Two noteworthy approaches exist for training on scaling. The first approach is *constrained scaling* (West, Ward, and Khosla, 2000). Constrained scaling attempts to calibrate the individual to the range of a scale. Typically, analysts are trained to match a physical stimulus (such as loudness) to a numeric scale (such as a 1-100 scale). Through multiple iterations across the full stimulus range, the individual receives feedback about the magnitude of the stimulus. The individual learns the use of the scale and is able to apply that scale to other domains. The result is significantly decreased variability in scaling results. Recent research (Boring, 2003b; West, Boring, and Moore, 2002) has shown that training on a physical stimulus such as loudness or brightness generalizes well to other dimensions and that constrained scaling can serve as a general-purpose method for increasing psychological scaling reliability. Nonetheless, the feasibility of using this method of scale calibration for expert elicitation has not yet been determined and warrants further investigation.

A second approach to preventing scaling biases is to provide training in probabilistic reasoning, specifically *Bayesian reasoning* (Sedlmeier and Gigerenzer, 2001). Extensive training is available on PSA, HRA, and specific expert elicitation methods. However, the importance of basic instruction in Bayesian reasoning and probabilistic techniques is commonly overlooked.

The key assumption here is that most individuals have some degree of probabilistic innumeracy and have not achieved expertise in assigning probabilities to events. Individuals tend to overweigh statistical base rates and neglect relevant information that would affect the event probability. Through training on Bayesian reasoning and the use of frequency information instead of raw probability scores, the authors argue, individuals are able to use the scale range available more accurately to reflect likelihood judgments.

These two examples—constrained scaling and Bayesian reasoning—illustrate the importance of giving the analyst explicit training on probabilistic scale use. Expert knowledge is benefited when it is coupled with honed elicitation. Training to calibrate likelihood estimation across analysts is an important step toward mitigating the effects of scaling biases on error and failure probabilities.

3.3 Univariate and Multivariate Judgments

An important concept in measurement in the physical sciences centers on the multidimensionality of measurement for any given object. As Kyburg (1984, p. 17) notes:

> Measurement is often characterized as the assignment of numbers to objects (or processes). Thus we may assign one number to a steel rod to reflect its length, another to indicate its mass, yet another to correspond to its electrical resistance, and so on. It is thus natural to view a quantity as a function whose domain is the set of things that quantity may characterize, and whose range is included in the set of real numbers.

Any object or event has a multitude of magnitude dimensions in which it may be measured. While in many cases these magnitude dimensions may be orthogonal, they are often interrelated.

The psychological analog to measurement in the physical sciences is scaling, including expert elicitation methods. Psychological scaling has univariate and multivariate components. If a single factor contributes to an event, then the judgment of that event is *univariate*. If a combination of factors contributes to an event, then the judgment of that event is *multivariate*.

Most reportable events in the safety-critical industries will feature a combination of contributing factors. This is, in part, due to deliberate safeguards and redundancies in operational processes, which minimize the chance that a single failure will escalate to become an off-normal event. For example, a maintenance electrician may accidentally reverse the polarity of a switch during installation. This error may be due to the electrician's fatigue at the end of the work shift. However, due to prescribed post maintenance checking, the electrician catches the error before it affects plant operations. If another contributing factor is added to the situation, the likelihood of the error leading to a reportable event increases. For example, if the procedures for installing the switch fail to specify post maintenance testing, the error may compound to affect operations adversely. The likelihood of error further increases as additional contributing factors are added. Fatigue and poor procedures might be joined by poor lighting at the switchboard where the switch is being installed. This poor lighting might make it difficult for the electrician to see the color of the wires being installed. Thus, ergonomics would escalate the probability of the error and the failure to correct it before it is put into service.

Table 1 depicts the considerations for using atomistic or holistic analysis for univariate and multivariate problems. In an atomistic approach to human reliability analysis, the analyst would classify each of the contributing factors to the event. The above example suggests that fitness for duty was low, that the switchboard featured poor maintenance ergonomics, and that there were issues with the procedures used for electrical maintenance. Performance shaping factors corresponding to these contributing factors would be flagged, and the human error probability would be computed accordingly. The performance shaping factor rubric or checklist forces the analyst to consider a variety of factors and therefore minimizes the chance of excluding important contributing factors. But, if the list of performance shaping factors in incomplete or fails to match the actual circumstances of the event, the analyst may overlook an important contributor or may need to use a supplemental holistic approach to model the complete circumstances of the event.

Table 1. Considerations for using atomistic or holistic judgment strategies for univariate and multivariate problem spaces for an event.

		Judgment Strategy	
		Atomistic	**Holistic**
Problem Space	**Univariate**	*Works well if one of the items on rubric/checklist matches the cause of the event.*	*Useful especially in unusual, previously unencountered situations. Works well if analyst avoids extraneous factors.*
	Multivariate	*Rubric/checklist helps analyst focus on relevant contributors to multivariate events.*	*Prone to inclusion of extraneous factors or scaling biases for multivariate events.*

In a holistic approach, the analyst synthesizes contributors to an event to determine the appropriate error probability. The open-ended nature of the holistic approach affords the analyst considerable flexibility in considering unusual contributing factors that may not be included in an atomistic checklist. Because the holistic approach does not necessarily provide guidance to zero in on common contributing factors, the holistic analyst is more likely than the atomistic analyst to consider factors extraneous to the event outcome.

Further, the holistic approach typically fails to provide clear guidelines for aggregating multivariate contributors to an event. Without a formal procedure, the aggregation of multivariate contributors may exhibit large inconsistencies within an individual analyst and across multiple analysts.

4 CONCLUSIONS

When dealing with the type of expert elicitation involved in PSA and HRA, expertise is not something that is readily manipulated. It is assumed that risk analysts possess a high degree of subject-matter expertise based on regulatory and operations experience as well as extensive formal training. However, it should be assumed that a subject-matter expert, without formal training in expert elicitation methods, is not necessarily an expert at making judgments about event likelihoods. By definition, the events

for which expert judgment is required occur so infrequently that the experts will not have extensive experience with them. Thus, the expert is called upon to generalize from other events for which there is sufficient theoretical and operating experience. These other events may or may not readily generalize, and different experts may use different types of events from their own experience to infer likelihoods. Moreover, expert elicitation methods offer considerably different methods for estimating event probabilities, and some methods, such as holistic approaches, offer little guidance on scaling event magnitudes and calculating event probabilities.

The manner in which experts make judgments can have a significant bearing on the quality of those judgments. It is the goal of this paper that future use of expert estimation as well as future method development in expert estimation will carefully consider scaling biases, the importance of training for expert estimation, and the differences between atomistic and holistic judgments and their consequences for univariate and multivariate problems. Without understanding the theoretic underpinnings of expertise and judgment processes—and accounting for them in practice and method development—expert elicitation risks not being an effective tool for PSA and HRA. The quality of estimation is potentially enhanced by consideration of these theoretical underpinnings and an active endeavor to improve expert elicitation processes for risk analysis through further exploration of these underpinnings.

5 REFERENCES

Anderson, J.R. 1995. *Cognitive Psychology and Its Implications, Fourth Edition*. New York: W.H. Freeman and Company.

Boring, R.L. 2003a. Human and computer-generated essay grades. *Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society*, 885-889.

Boring, R.L. 2003b. Improving human scaling reliability. *Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society*, 1820-1824.

Comer, M.K., Seaver, D.A., Stillwell, W.G., & Gaddy, C.D. 1984. *Generating Human Reliability Estimates Using Expert Elicitation, Volume 1. Main Report*, *NUREG/CR-3688.* Washingtion, DC: US Nuclear Regulatory Commission.

Ericsson, K.A., Krampe, R.T., & Tesch-Römer, C. 1993. The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*: 363-406.

Ettenson, R., Shanteau, J., & Krogstad, J. 1987. Expert judgment: Is more information better? *Psychological Reports*, *60*: 227-238.

Gertman, D., Blackman, H., Marble, J., Byers, J., Haney, L., & Smith, C. 2005. *The SPAR-H Human Reliability Analysis Method. NUREG/CR-6883*. Washington, DC: US Nuclear Regulatory Commission.

Hammond, K.R., Hamm, R.M., Grassia, J., & Pearson, T. 1987. Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE Transactions on Systems, Man, and Cybernetics*, *17*: 753-770.

Helson, H. 1964. *Adaptation-Level Theory*. New York: Harper & Row.

Köhler, W. 1947. *Gestalt Psychology. Revised Edition.* New York: Liveright.

Kyburg, H.E. 1984. *Theory and Measurement*. Cambridge, UK: Cambridge University Press.

Madigan, R.J., & Brosamer, J.J. 1991. Holistic grading of written work in introductory psychology: Reliability, validity, and efficiency. *Teaching of Psychology*, *18*: 91-94.

Meyer, M. A., & Booker, J. M. 1990. *Eliciting and Analyzing Expert Judgment, A Practical Guide*, *NUREG/CR-5424*. Washington, DC: US Nuclear Regulatory Commission.

Poulton, E.C. 1989. *Bias in Quantifying Judgments.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Pylyshyn, Z.W. 1984. *Computation and Cognition. Toward a Foundation for Cognitive Science*. Cambridge, MA: MIT Press.

Rasmussen, J. 1983. Skills, rules, knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man and Cybernetics*, 13: 257-266.

Seaver, D.A., & Stillwell, W.G. 1983. *Procedures for Using Expert Judgment to Estimate Human Error Probabilities in Nuclear Power Plant Operations*, *NUREG/CR-2743.* Washington, DC: US Nuclear Regulatory Commission.

Sedlmeier, P., & Gigerenzer, G. 2001. Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130*: 380-400.

Simon, H.A., & Chase, W.G. 1973. Skill in chess. *American Scientist*, *61*: 394-403.

Stevens, S.S. 1975. *Psychophysics. Introduction to Its Perceptual, Neural, and Social Prospects.* New York: Wiley.

Stillwell, W.G., Seaver, D.A., & Schwartz, J.P. 1982. *Expert Estimation of Human Error Probabilities in Nuclear Power Plant Operations: A Review of Probability Assessment and Scaling*, *NUREG/CR-2255*. Washington, DC: US Nuclear Regulatory Commission.

Titchener, E.B. 1911. *A Text-Book of Psychology*. New York: Macmillan.

Weiss, D.J., & Shanteau, J. 2003. Empirical assessment of expertise. *Human Factors*, *45*: 104-114.

West, R.L., Boring, R.L., & Moore, S. 2002. Computer augmented psychophysical scaling. *Conference Proceedings of the 24th Annual Cognitive Science Society,* 932-937.

West, R.L., Ward, L.M., & Khosla, R. 2000. Constrained scaling: The effect of learned psychophysical scales on idiosyncratic response bias. *Perception & Psychophysics*, *62*: 137-151.

Wilson, J.M. 1994. Network representations of knowledge about chemical equilibrium: Variations with achievement. *Journal of Research in Science Teaching*, *31*: 1133-1147.