ORGANIZATIONAL BEHAVIOR AND HUMAN DECISION PROCESSES
Vol. 74, No. 2, May, pp. 89–117, 1998
ARTICLE NO. OB982771

Cross-Cultural Variations in Probability Judgment Accuracy: Beyond General Knowledge Overconfidence?

J. Frank Yates

The University of Michigan

Ju-Whei Lee

Chung Yuan University, Republic of China

Hiromi Shinotsuka

Hokkaido University, Sapporo, Japan

and

Andrea L. Patalano and Winston R. Sieck

The University of Michigan

Previous studies have revealed surprising and persistent cross-cultural variations in overconfidence, whereby respondents in some Asian cultures (e.g., Chinese) exhibit markedly higher degrees of overconfidence than respondents in other cultures (e.g., in the United States and Japan). Most of those demonstrations have entailed general knowledge tasks (e.g., answering

This research was supported by U.S. National Science Foundation Grant SES92-10027 to the University of Michigan, Grant NSC84-2413-H033-002 from the R.O.C. National Science Council to Chung Yuan University, and by the Psychology Department at Hokkaido University. We are especially grateful to Kevin Biolsi for his assistance in programming the experiments described here and to Jonathan Emmett, LeAnn Franke, and Joe Magee for their help in collecting and processing the data. We also appreciate the useful discussions we have had about this research with other members of the Michigan Culture and Cognition Program as well as the comments on a previous version of this article provided by Ido Erev and an anonymous reviewer.

Correspondence and reprint requests should be addressed to: J. Frank Yates, Judgment and Decision Laboratory, Department of Psychology, University of Michigan, 525 East University Avenue, Ann Arbor, MI 48109-1109, e-mail: jfyates@umich.edu; Ju-Whei Lee, Department of Psychology, Chung Yuan University, Chungli, Taiwan, ROC. e-mail: jwlee@cchp01.cc.cycu.edu.tw; or Hiromi Shinotsuka, Department of Psychology, Hokkaido University, Sapporo 060, Japan, e-mail: shino@hubs.hokudai.ac.jp.

0749-5978/98 $25.00

questions such as whether Europe is larger than Australia). The present studies sought to determine whether such cross-cultural variations extend to judgments about the kinds of events that bear upon more common practical decisions and to aspects of accuracy other than overconfidence. Subjects in Taiwan, Japan, and the United States made probabilistic differential diagnoses of fictional diseases in a stochastic artificial ecology. Results revealed that previously observed cross-cultural variations do indeed generalize. The data were also informative about several potential accounts for such variations, e.g., arguing against a proposal that they rest on different emphases on discrimination rather than calibration, but consistent with the influences of culture-specific cognitive customs, including responsiveness to explicitly displayed information, regardless of its presumed validity.

Imagine the following business conversation:

A: "If the chances are better than 75% that they'll deliver on time, we ought to go with Consolidated. What *are* the chances?

B: "Oh, I'd say about 80%."

A: "OK. Then let's do it."

Now consider this question, which captures the primary practical aims of the research described here: Should the nationalities or "cultures" (in the broad sense) of A and B matter to the quality of the decision they are making? If the answer to this question is "Yes," then other questions follow, such as how and why the cultures matter.

The scenario described above is an instantiation of the "threshold" approach to choice formalized in technologies such as decision analysis (see, for example, Clemen, 1991). However (in the West, at least), the basic idea embodied in that approach is quite general and intuitively compelling, and it is also thought to capture what occurs in many practical decision situations. Thus, suppose a given choice alternative promises favorable consequences (e.g., high profits) provided that a given event occurs (e.g., timely delivery). Then the attractiveness of that option should increase in relation to the chances of that critical event. Indeed, implicitly if not explicitly, there must be some minimal degree of certainty in that event's occurrence (e.g., a 75% threshold) such that beyond that point the option is so appealing that the decision maker feels obliged to pursue it.

Now, the quality of decisions reached via such approaches depends on many things. Nevertheless, it is clear that those decisions can turn out no better than the accuracy of the degrees of certainty—typically real people's likelihood judgments—at the core of the means by which those choices are made. Suppose an executive routinely believes that laggard bidders have strong chances of performing well and that good ones have weak chances and that she chooses contractors on the basis of those beliefs. Then that executive is bound to lose money. Recognition of such possibilities in all manner of situations provides the

pragmatic motivation for the tremendous amount of attention that judgment research has experienced over the years: Just how accurate *are* people's judgments? How can high degrees of accuracy be assured?

Questions of culture, like those posed above, are simply special cases of concerns like these. Interest in them has been fanned by the fact that, over the past two decades, there have been numerous and consistent demonstrations of cross-cultural variations in probability judgments about general knowledge (e.g., Lee, Yates, Shinotsuka, Singh, Onglatco, Yen, Gupta, & Bhatnagar, 1995; Whitcomb, Onkal, Curley, & Benson, 1995; Wright, Phillips, Whalley, Choo, Ng, Tan, & Wisudha, 1978; Yates, Zhu, Ronis, Wang, Shinotsuka, & Toda, 1989). The following example illustrates the prototypical task used in such studies. The subject is first asked: "For which is the gestation period longer: (a) humans or (b) chimpanzees?" After picking an alternative, the subject then reports a probability judgment between 50 and 100% that the selected answer is indeed correct. Probability judgments are said to be well-calibrated to the degree that those judgments match the relative frequencies with which the pertinent target events actually occur (e.g., events assigned 70% judgments really happen about 70% of the time). Usually (although not always), people's probability judgments about their general knowledge are miscalibrated in a particular way. On average, they are higher than the proportions of questions respondents actually answer correctly, a phenomenon commonly described as "overconfidence." It comes as a surprise to most people (Yates, Lee, & Shinotsuka, 1996) that such overconfidence is typically greater for subjects in Asian cultures than for those in the West. Responses of subjects in Japan and Singapore provide notable exceptions to this pattern.

Suppose that the kinds of cross-cultural judgment differences found in general knowledge research extend to the types of practical situations exemplified by our fictional business conversation. Then the decisions predicated on those judgments, as well as their quality, should be affected correspondingly. In one version of the scenario described, suppose that decision maker B is an American who reports a probability judgment of 70%, implying that no contract is extended to Consolidated. In another version of the scenario, decision maker B is a Chinese who has access to the same information as the American decision maker B. We should not be surprised to see this new decision maker B report a more extreme 80% degree of certainty in timely delivery, thereby leading to a *favorable* decision for Consolidated, a decision that has a good chance of turning out badly. Should we, in fact, expect the kind of generalization assumed in the example? On theoretical grounds, the answer to this question is unclear. It depends on the extent to which judgments about general knowledge and about the kinds of events bearing on typical practical decisions rest on the same mechanisms. A definitive assessment has yet to be made, but several authors (e.g., Wright & Ayton, 1986) have argued against simply assuming the generalizability to other contexts of conclusions established in general knowledge studies.

The best way to assess generality is empirically, and there have been several attempts to apply that strategy in the case of cross-cultural overconfidence

comparisons. In separate studies, Wright and Wisudha (1982) asked British and Indonesian subjects to make probability judgments for a potpourri of future events. A sample item seen by the British subjects was this: "At least one national leader (president or prime minister, etc.) (a) will or (b) will not die during the next 30 days." The following was an Indonesian example: "When will the Cengkareng airport be operational? (a) before the end of 1978, (b) after the end of 1978." Wright and Wisudha found that the calibration of their Indonesian subjects' future-event judgments was much better than the calibration of general knowledge judgments reported in previous research with Indonesian respondents. There were, nevertheless, marked British–Indonesian differences in the calibration of the future-event judgments, with the Indonesian judgments being largely overconfident and the British judgments underconfident. Yates *et al.* (1989) asked their American and mainland Chinese subjects to make probability judgments concerning the future values of various quantities, e.g., the next-day high temperature in a designated city in the United States or China, respectively. Consistent with the general knowledge results, the Chinese subjects were decidedly more overconfident than the Americans. Zhang (1992) examined the calibration of probabilistic predictions of various economic indicators made by professional forecasters in Beijing. He found the same kind of extreme Chinese overconfidence observed by Yates *et al.* (1989).

Results like these support the expectation that previously documented cross-cultural variations in general knowledge probability judgments do, in fact, apply to the kinds of judgments that drive common decisions. Yet, in every case, there are reasons for caution in accepting this conclusion. For example, in none of the studies did subjects in different locations consider the very same events and with the benefit of identical current information sources. (In even the "tightest" cross-cultural comparisons of probability judgment, there has been no control for participants' prior learning.) And in some instances (e.g., Wright & Wisudha, 1982), there were not direct statistical comparisons made between judgments offered by subjects in the different countries studied. Then there is the nature of the events considered. Miscellaneous, unrelated future events (e.g., national leaders dying and airports opening) are important, since they do indeed support some important day-to-day decisions. Yet, as discussed below, studying judgments about such events poses more difficult analytic problems than studying repeatable events. By "repeatable" events we mean those that, on substantive grounds, are highly similar from one instance to the next, e.g., rises in the prices of various stocks, incidences of pneumonia in different patients, or timely deliveries by competing contractors.

So, the first primary aim of the present research was to address the generalizability question rigorously, with due attention to the kinds of control and analytic issues that clouded previous efforts. Simply put, the question was this: When making probability judgments about repeatable events that support common practical decisions, do people exhibit the same cross-cultural variations in overconfidence they display in their general knowledge judgments? In the process of answering this culture question, the present studies should also enlighten us about whether the overconfidence observed in general knowledge

judgments should be expected in judgments reported by *any* respondent group about other kinds of events.

Calibration, including overconfidence, is unquestionably important, especially when decisions are made via procedures such as decision analysis, e.g., with threshold rules that prescribe different choices when probabilities fall within different ranges along the continuum. It is thus understandable that calibration has been subjected to enormous scrutiny. Nevertheless, calibration is far from the whole story about accuracy. Defensible conceptions of overall probability judgment accuracy acknowledge that there is more to the construct besides the notion embodied in calibration, i.e., the match between the numerical character of probability statements and the relative frequencies of events assigned those statements. (What, for instance, happens on the individual occasions that are "hidden" in a relative frequency, the way all individual case variation is obscured in any kind of average?) That is why contemporary overall accuracy measures go beyond that narrow characterization (cf. Yates, 1990, chapter, 3, 1994). This recognition also leads to important questions of cross-cultural variations in overall probability judgment accuracy, e.g., whether they exist, what their nature might be, and how they can be explained. To the best of our knowledge, Yates *et al*. (1989) were the first to seek answers to such questions. And those authors found, surprisingly, that the overall accuracy levels of their mainland Chinese, Japanese, and American subjects' general knowledge probability judgments were virtually the same. Hence, a second key issue addressed in the present research was whether this conclusion would hold for judgments about the kinds of events that underlie typical practical decisions.

Several authors have shown how common measures of overall probability judgment accuracy can be partitioned or decomposed into submeasures reflecting meaningfully distinct and imperfectly correlated elements of overall accuracy (see Yates, 1982, 1990, 1994, for reviews). Calibration is only one of those dimensions, and quite arguably not the most significant one, depending on how the judgments are used in actual decision making. A particularly noteworthy dimension other than calibration is discrimination, or resolution, as it is sometimes called. We have noted that calibration reflects a person's ability to correctly apply numerical labels to his or her degrees of belief in an event's occurrence, where "correctly" implies matching with relative frequencies. In contrast, discrimination reflects the person's tendency to say *different* things—irrespective of their numerical labels—on occasions when the target event (e.g., timely delivery) occurs as opposed to when it does not. Put another way, discrimination refers to the strength of *any* form of statistical association between a person's judgments and the events of interest (e.g., as does a chi square statistic computed for an ordinary contingency table; cf. Yaniv, Yates, & Smith, 1992).

An illustration: Consider a person who, for simplicity of discussion, reports only three different levels of certainty—30, 50, and 90%—in a situation where a target event occurs 60% of the time overall. At one extreme, the person's judgments exhibit no discrimination at all if the target event is observed about 60% of the time regardless of the judgment the person happens to report, e.g.,

it is 60% when she reports 30, 50, or 90%. At the other extreme, the person's judgments are perfectly discriminative if there is never a pair of instances such that the same judgment is assigned on both occasions but in one of them the target event occurs whereas in the other it does not. One (intentionally bizarre) form that perfect discrimination could take (among many) would entail the target event always occurring whenever the person reports 30% judgments and never happening on those occasions when she announces 50 or 90% judgments (see Yates, 1990, chapter 3, for further discussion).

A good case can be made that, especially in some practical domains, discrimination is a more important judgment quality than calibration (Yaniv *et al.*, 1991; Yates, 1982). Why? Good calibration often can be achieved rather easily post hoc, via mere mathematical transformations of a given set of judgments. Good discrimination cannot be attained so readily. This quality requires that a judge have access to information or "cues" that have strong associations with the pertinent target event, access that is often difficult and sometimes impossible to acquire. The judge must also understand the form and strength of those associations and how to exploit them (cf. Yates, 1994). Thus, a judge who somehow demonstrates outstanding discrimination performance is a distinctively valuable resource.

These considerations are why one finding reported by Yates *et al.* (1989) was especially noteworthy and therefore provided additional impetus for the present studies. Those investigators discovered that, although their Chinese subjects' general knowledge judgments exhibited worse calibration than those of their American counterparts, their discrimination was better (and thereby allowed for the equivalent overall accuracy that was observed). Thus, our third key objective was to determine whether similar outstanding Chinese discrimination is evident in nongeneral knowledge judgments.[1]

Before proceeding, we must address an analytic issue that bears importantly on measures of probability judgment accuracy, discrimination in particular. Despite its criticality, the issue is almost never acknowledged in the literature, perhaps because it is so subtle. Target events in accuracy analyses are sometimes defined internally but at other times externally, where the referent for these descriptors is the perspective of the person making the assessments, the "judge." In internal definition, the target event assumes a form such as the following: A = "My chosen alternative is in fact correct" or A = "The event I categorically predict to occur really will occur." Thus, the person has made some deterministic judgment and now the question is the adequacy of that prior statement. Implicitly or explicitly, all analyses of overconfidence presume internal definition. For instance, for a typical general knowledge item such as our earlier illustration, the subject reports a 50–100% probability judgment for the target event A = "My choice of alternative (a) or (b) as the correct answer really was correct." Overconfidence is inferred when the average of such judgments exceeds the proportion of chosen alternatives that actually

[1] Note that superior Chinese discrimination has not always been found even for general knowledge judgments (e.g., Yates, Lee, Levi, & Curley, 1990).

were correct. The label "internal" applies because the pertinent probability judgments refer to the person's own prior judgments (i.e., selections of alternatives). In external definition, the target event makes no reference to any such prior categorical assessment by the judge. In our business scenario, the target event could be represented as A = "The contractor will deliver on time," in weather forecasting, it might be A = "Precipitation will occur," and in a legal trial, it could be A = "We will win the case."

Recognize that, in some situations, the accuracy analyst has a choice about how to define the target event whereas in others there is no (meaningful) choice. In the examples just described, there is an option. Thus, for external target event A = "The contractor will deliver on time," we could also define a corresponding internal target event A* = "My categorical prediction that the contractor either will or will not deliver on time will be borne out." Transformations between judgments for equivalent internally and externally defined target events in such cases are predicated on an assumption of additivity for probability judgments for complementary events, i.e., $P'(E) + P'(E^c) = 1$, for arbitrary event E. When he actually renders his assessments, the judge might do so via a two-stage judgment routine. In the first stage, the judge literally makes a categorical judgment and then in the second stage reports a probability judgment about the correctness of that previous selection. For example, the judge might predict categorically that "Consolidated will fail to deliver on time" and then indicate 70% certainty that that prediction would prove correct. If the original internally defined target event A = "My prediction will be correct" is redefined externally as A* = "Will deliver on time," then the inferred probability judgment for A* would be 100% − 70% = 30%. Now, if the judge instead engages in a single-stage judgment routine, he would report a 0–100% probability for a specific externally defined target event (e.g., A = "Will deliver on time"). And if (in a dichotomous-event situation) that judgment happens to be above 50% (e.g., 85%), it is inferred that, in a two-stage routine, the judge would have predicted categorically the event specified explicitly in the target event (e.g., "I predict on-time delivery") and a probability of being correct that is the same as that assigned to the externally defined event (e.g., 85%). If that judgment is below 50%, the inference is a categorical selection of the complementary event and a probability judgment of correctness that is 100% less the probability judgment reported for the externally defined event.

When the various cases considered by a judge are highly heterogenous and unrelated to one another, external definition can still be done, but its interpretation is suspect. That is why, in such circumstances, analyses are virtually always performed for internally defined target events. General knowledge studies involving questions having no principled connection with one another provide a good example. One possible externally defined target event might be A = "The alternative on the left is correct." But if (as is typically the case), experimenters assign correct alternatives to the left and right display positions randomly, then any analyses focusing on that target event would be uninformative. It is hence no surprise that in such studies (as well as in forecasting studies such as that of Wright & Wisudha, 1978) all analyses are carried

out with internally defined target events, e.g., A = "My selection/prediction was correct."

Internal definition is perfectly sensible for studying overconfidence, but it is problematic for examining accuracy dimensions such as discrimination. That is because discrimination measures are difficult to interpret when targets are defined internally (cf. Yates, 1982). Discrimination skill is typically measured by the following discrimination index (DI) due to Murphy (1973; see also Yates, 1990, 1994):

$$\mathrm{DI} = (1/N) \Sigma\ N_j(\bar{d}_j - \bar{d})^2. \tag{1}$$

In this expression, j indexes the judgment response category f_j, e.g., $f_1 = 0\%$, $f_2 = 10\%$, $f_3 = 20\%$, etc., and the summation ranges over all categories. $\bar{d}$ is the overall base rate for the target event, represented as the mean of an indicator variable d called the outcome index, which takes on the value 1 when the target event occurs and 0 otherwise. $\bar{d}_j$ is the conditional base rate, the relative frequency of target event occurrences given that the judgment response category f_j is offered. For example, suppose the judge reports judgment $f_3 = 20\%$ on $N_3 = 50$ occasions and that the target event actually occurs on 15 of those occasions. Then $\bar{d}_3$ would be 15/50 or 30%. Of course, N is simply the total number of judgments rendered altogether, over all categories. It is straightforward to show that the maximum (and best) value of DI is the variance of d, i.e., $\mathrm{DI}_{\mathrm{Max}} = \mathrm{Var}\ (d) = \bar{d}\ (1 - \bar{d})$. Now, suppose the target event is defined internally, say, A = "My chosen answer is correct," implying, of course, that A^c = "My chosen answer is incorrect." It is apparent how $\mathrm{DI}_{\mathrm{Max}}$ becomes rather slippery, depending as it does on $\bar{d}$, the proportion of correct answer selections, which can (and usually does) vary considerably from one person to the next.

An even more troublesome interpretation problem is illustrated by the following: Suppose a judge has perfect discrimination ability for an internally defined target event, A = "My chosen alternative is correct," in the interests of concreteness. Recall that this means that the judge reports different probability judgments (say, "70%," for the sake of simplicity) on those occasions when the target event is going to occur (i.e., his chosen answer will prove to be correct) from those he reports (say, "40%") on those occasions when the target event is *not* going to occur (i.e., his chosen answer will prove to be wrong). It seems inevitable that, if such a situation arose, the judge would change his mind in every case where his initially chosen answer was going to be in error. The result would be $\bar{d} = 1$, perfect performance, and the discrimination index would become degenerate. For present, analytic purposes, the problem is how to conceptualize the ideal value of DI when the target event is internally defined, as in the Yates *et al.* (1989) study, where Chinese subjects' general knowledge judgments exhibited better discrimination than their American subjects' judgments. And hence our aim in the present research is to determine whether similar, easy-to-interpret cross-cultural differences in discrimination occur when target events are defined externally in a meaningful way.

Our final objectives in the present studies were to shed light on possible explanations for cross-national variations in probability judgment accuracy, overconfidence in particular. Three potential accounts had special interest. The first suggests that groups who exhibit weak calibration, including markedly high degrees of overconfidence, do so because of "relative emphasis." Specifically, for any number of plausible reasons, their cultures might attach greater significance to discrimination than calibration (cf. Yates *et al.,* 1989). If so, this would allow for stronger discrimination than calibration (although admittedly it would not provide a positive account for the resulting miscalibration taking the form of overconfidence instead of, say, underconfidence). At least minimal support for this proposition would be provided if we found evidence of strong discrimination for subjects belonging to a group with a high degree of overconfidence.

The second proposed explanation for cross-cultural variations in overconfidence arises from recent suggestions that (apparent) overconfidence might be at least partly a reflection of the inconsistency with which people report their true beliefs (e.g., Erev, Wallsten, & Budescu, 1994). In the present context, the suggestion would be that cross-cultural variations in overconfidence are mediated by consistency variations. Support for this idea would be revealed in the present research if there were strong assocations between overconfidence and reliability measures as well as particular patterns of cross-cultural differences in those measures.

The third class of possible accounts for cultural differences in overconfidence focuses on cognitive customs, including learning strategies. Some work, for instance, has suggested especially strong Chinese premiums on memory (e.g., Liu, 1986) that would exacerbate overconfidence in some circumstances (see, for example, arguments by Fischhoff, Slovic, & Lichtenstein, 1977). Previous studies (e.g., Yates, Lee, & Shinotsuka, 1992) have suggested that Chinese modes of generating arguments should support overconfidence as well. The present studies were designed to be informative about propositions like these.

Our approach entailed a simulation, a controlled yet reasonably realistic artificial world or "ecology." In this ecology, a designated target event was statistically related to, and hence (imperfectly) predictable from, a collection of facts or "cues" about the given cases. Subjects in the participating countries would then be immersed in this ecology, with all subjects starting from the same point—ignorance. The subjects would render probability judgments for the designated target event and be given case-by-case feedback about the actual occurrences and nonoccurrences of that event, as in standard multiple-cue probability learning studies, but with some key differences. Thus, over time (and as in real life), the subject could learn to improve the quality of his or her judgments via induction of the properties of the ecology.

We conducted two studies. In Study 1, on each trial every subject was exposed to all the available cues and, after making a judgment, was also told about the target event's occurrence or nonoccurrence. In many real-life situations, people are free to choose which information they will examine as a means of arriving at their judgments. To simulate such circumstances and to pursue

specific questions unaddressable with fixed displays, in Study 2 the subject was allowed to choose as many or as few of the available cues as desired prior to offering a judgment.

STUDY 1: FIXED SYMPTOM DISPLAYS

Method

Subjects

The volunteers who participated in the study were psychology students in three countries where previous studies have documented consistent cross-cultural variations in overconfidence: 56 at Chung Yuan University in Taiwan, 41 at Hokkaido University in Japan, and 46 at the University of Michigan in the United States.

Cover Story and Ecology

The subject was asked to assume the role of a physician in the following scenario: Two new (and fictitious) diseases have appeared in the physician's community, "Trebitis" and "Philiosis." It has been established that Trebitis and Philiosis patients share the following symptoms: sore throat, nose bleeds, loss of appetite, dizziness, and muscle aches. During the subject's medical career in the laboratory, the subject will see a series of patients, each showing all these symptoms, therefore indicating that the patient has either Trebitis or Philiosis. Each patient also either does or does not have each of the following six symptoms, which the subject and some colleagues suspect *might* be helpful in distinguishing Trebitis from Philiosis: nausea, earache, coughing, rash, irregular heartbeat, and stiff joints. Thus, for each case, the task is to arrive at a probabilistic differential diagnosis between Trebitis and Philiosis.

The ecology was designed to be challenging and realistic yet "learnable" and hence not overly frustrating to subjects. The following are key features of the ecology:

Base rates. The base rate for Trebitis was .60 and hence that for Philiosis was .40. Each of the symptoms had a base rate of .50.

Validities. Following are the validities of the available symptoms for distinguishing Trebitis from Philiosis: nausea (N): .61 (.60); earache (E): .41 (.40); coughing (c): .41 (.40); rash (R): .20 (.20); irregular heartbeat (H): .20 (.20); and stiff joints (J): .00 (.00). The validities are indicated in two ways here. The first index listed is an ordinary Pearson product–moment correlation coefficient, r (with Trebitis coded 1 and Philiosis 0 and the presence and absence of a given symptom coded 1 and 0, respectively). The second index, in parentheses, is a "contingency statistic" commonly used in contingency judgment research, ΔP. In this instance, for a given symptom, ΔP is defined as follows:

$$\Delta P = P(\text{Trebitis}|\text{Symptom Present}) - P(\text{Trebitis}|\text{Symptom Absent}). \quad (2)$$

As is apparent from the listing, the ecology was designed to include cues with a range of validities, from high to nil. This was intended to mimic the reality that people are often confronted with information that varies in its actual usefulness, including facts that have no value at all.

Predictability. Altogether, subjects saw a total of 120 "patients," divided randomly into two blocks of 60. The best-fitting ordinary regression models for predicting disease status from the symptoms yielded multiple correlations of $R = .74$ ($R^2 = .5476$) on Block 1 and $R = .75$ ($R^2 = .5625$) on Block 2. Thus, the two blocks were clearly comparable with each other. Tape, Heckerling, Ornato, and Wigton (1991) studied Illinois, Virginia, and Nebraska physicians' probabilistic diagnoses of pneumonia. They also constructed regression models of patients' actual pneumonia conditions in terms of various symptoms and patient characteristics. The multiple correlations for the best models in Illinois, Virginia, and Nebraska were, respectively, .39, .64, and .59. Hence, the artificial ecology here was somewhat "easier" than the real ones studied by Tape et al., but this is not unreasonable given that the present subjects were laypersons and were working in constrained circumstances.

An important alternative perspective on the predictability of disease status in the ecology is provided by the quadratic scoring rule sometimes called the probability score (cf. Yates, 1990):

$$\mathrm{PS} = (f - d)^2. \tag{3}$$

Here, $f = P'(\mathrm{A})$ is the probability judgment for the target event A (e.g., "This patient has Trebitis"). As in our discussion of discrimination measures, d is the outcome index, which assumes the value 1 when A occurs and 0 otherwise. The most frequently used measure of overall probability judgment accuracy is the mean of PS over a given sample of cases, often described as the Brier (1950) score:

$$\overline{\mathrm{PS}} = (1/N)\,\Sigma\,(f - d)^2. \tag{4}$$

$\overline{\mathrm{PS}}$ ranges between 0 and 1, where smaller scores are better, with 0 being ideal. A commonly discussed standard of comparison is the score $\overline{\mathrm{PS}} = .25$ earned by a "uniform judge" who always says that the target event and its complement are equally likely, i.e., $f = .5$. Another is the score $\overline{\mathrm{PS}} = \bar{d}\,(1 - \bar{d})$ earned by the "sample base rate judge" who, for every case, reports the sample base rate as a judgment, i.e., $f = \bar{d}$; here such a judge would achieve $\overline{\mathrm{PS}} = .24$ since $\bar{d} = .6$.

Logistic regression models were created for the blocks of patients in the present ecology. For Block 1, the model was

$$p = e^Y/[1 + e^Y], \tag{5}$$

where $p = P^*(\mathrm{A})$ is the model's probability "judgment" for the target event and

Y was the following linear combination of indicator variables for the presence and absence of the symptoms:

$$Y = -4.56 + 3.97N + 1.89E + 2.29C + 1.21R + .71H + .99J \qquad (6)$$

The model derived for Block 2 was similar. The Block 1 model achieved a value of $\overline{PS} = .1004$ on the Block 1 cases and .1040 when cross-validated on the Block 2 cases. The values of $\overline{PS}$ earned by the Block 2 model on the cases in Blocks 2 and 1, respectively, were .1008 and .1057. A point of reference is provided by Levi's (1986) study of coronary artery disease diagnoses. Levi's best model achieved a value of $\overline{PS} = .1480$ in an ecology where the base rate for the disease was $\bar{d} = .66$. Thus, once again, the present ecology appears to have been relatively easy yet not dramatically so.

Procedure

The procedure was conducted almost entirely via microcomputer, in the subject's native language, with all materials having been translated and back-translated using standard procedures (cf. Brislin, 1970). The program introduced the scenario to the subject as "physician." The introduction emphasized several points, including: (a) the need to learn, over time, to make good diagnoses; (b) that each symptom may or may not be useful in distinguishing Trebitis from Philiosis; (c) what it means for a symptom to have high, medium, low, or nil diagnostic power; and (d) that the diagnostic process is inherently probabilistic rather than deterministic, unlike what the subject might have been accustomed to seeing in other psychology experiments.

On each trial, the subject: (a) was presented with a new patient and that patient's complete symptom profile; (b) indicated whether it was more likely that the patient had Trebitis or Philiosis; (c) reported a probability between 50 and 100% that that categorical diagnosis was correct; (d) received feedback about what was "eventually determined" to be the patient's actual condition; and (e) received an accuracy score indicating the "quality" of the reported diagnosis. The program explained carefully how probability judgments and accuracy scores should be interpreted. Accuracy scores were linear transformations of the probability score (PS) such that higher scores indicated greater accuracy; explicitly, Score = 300(1 − PS). The program emphasized that the scoring procedure had a special characteristic technically called "properness" (a term not used with the subjects), which implied that it was in the subject's interests to be perfectly candid in reporting his or her true judgments (cf. Yates, 1990, chapter 8). The subject also learned that, out of each sequence of five subjects in the experiment, the one with the best average accuracy score would receive a bonus payment of $10 (or its equivalent in Taiwan and Japan). This was intended to encourage effort and accuracy and was described as analogous to developing a good reputation and a successful real-life medical practice. The subject made diagnoses for the 60 Block 1 patients during the initial session, which lasted no more than 50 min. The subject made assessments for the

remaining block of 60 patients in a second session scheduled exactly one week later.

Results and Discussion

The various effects described below were tested via both parametric and nonparametric methods since the sampling distributions for some of the statistics have not been established. The conclusions indicated by those analyses were virtually identical. To conserve space, here we present only the results of the parametric tests.

Overall Accuracy

Panel (a) of Fig. 1 shows the mean values of $\overline{PS}$ earned by subjects in each country and for each block of 60 patients seen in Sessions 1 and 2, when the subjects presumably were developing their judgment strategies and then applying them in mature form, respectively.[2] The first thing to notice in Fig. 1 is how the subjects' performance compared with that of the uniform judge (i.e., $\overline{PS} = .25$), who would have reported, for each and every patient, a probability of 50% that the patient had Trebitis, something of a "minimalist" standard. Observe that only the Japanese subjects met that norm initially. Learning clearly occurred in that $\overline{PS}$ improved significantly from Session 1 to Session 2, $F(1, 280) = 19.21$, $p < .001$. Indeed, in Session 2, only the Taiwanese subjects failed, on average, to outperform the uniform judge. But also note that none of the subject groups ever approached the standard of the logistic regression model ($\overline{PS} \approx .10$). That is, in principle, the subjects could have performed *much* better than they actually did.

Figure 1 further makes apparent the nature of the substantial country effect on $\overline{PS}$ that occurred, $F(2, 280) = 31.69$, $p < .001$. Although there were no significant differences in the overall accuracy levels of the Japanese (JPN) and American (USA) subjects' judgments, each of these was higher than that of the judgments expressed by the Chinese in Taiwan (TWN), $t(95) = 6.18$ and 3.99, $p < .001$, for TWN vs JPN in Sessions 1 and 2, respectively; $t(100) = 5.05$, $p < .001$, and $t(100) = 2.72$, $p < .01$, for TWN vs USA in Sessions 1 and 2, respectively.

Overconfidence

Overconfidence is typically indexed by a bias statistic defined on judgments relative to internally defined target events, in this case, $f = P'(A)$, where A = "I was correct in my selection of Trebitis or Philiosis as the patient's actual disease." Specifically, if the outcome index d is such that $d = 1$ if A occurs and 0 otherwise, then the overconfidence measure is

[2] We can show mathematically that, interestingly, for any given set of data, $\overline{PS}$ necessarily is unaffected by whether the target event is defined internally, e.g., "I am correct in my selection of Trebitis or Philiosis as the patient's actual disease," or externally, e.g., "This patient's actual disease is Trebitis." Thus, the definitional distinction is irrelevant here.

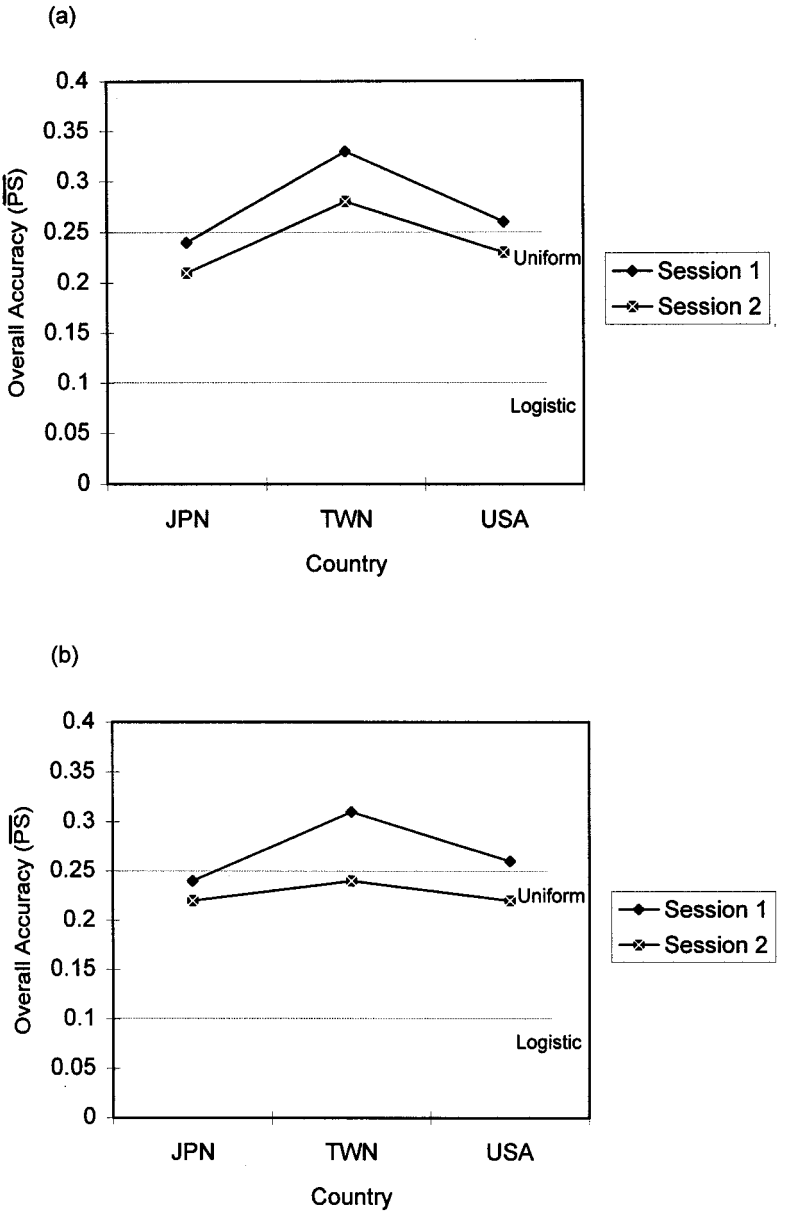

FIG. 1. Mean measures of overall accuracy ($\overline{PS}$) by country and session: (a) Study 1 (fixed symptom display), (b) Study 2 (discretionary symptom display). Notes: Smaller values of PS imply greater accuracy. JPN, Japan; TWN, Taiwan; USA, United States.

$$\text{Bias} = \bar{f} - \bar{d} \tag{7}$$

for which positive values indicate overconfidence and negative ones underconfidence. As Panel (a) in Fig. 2 shows, all three subject groups were overconfident in both sessions of Study 1. In fact, the observed levels of overconfidence were markedly higher than those usually found in general knowledge studies (e.g., Yates *et al.*, 1989).

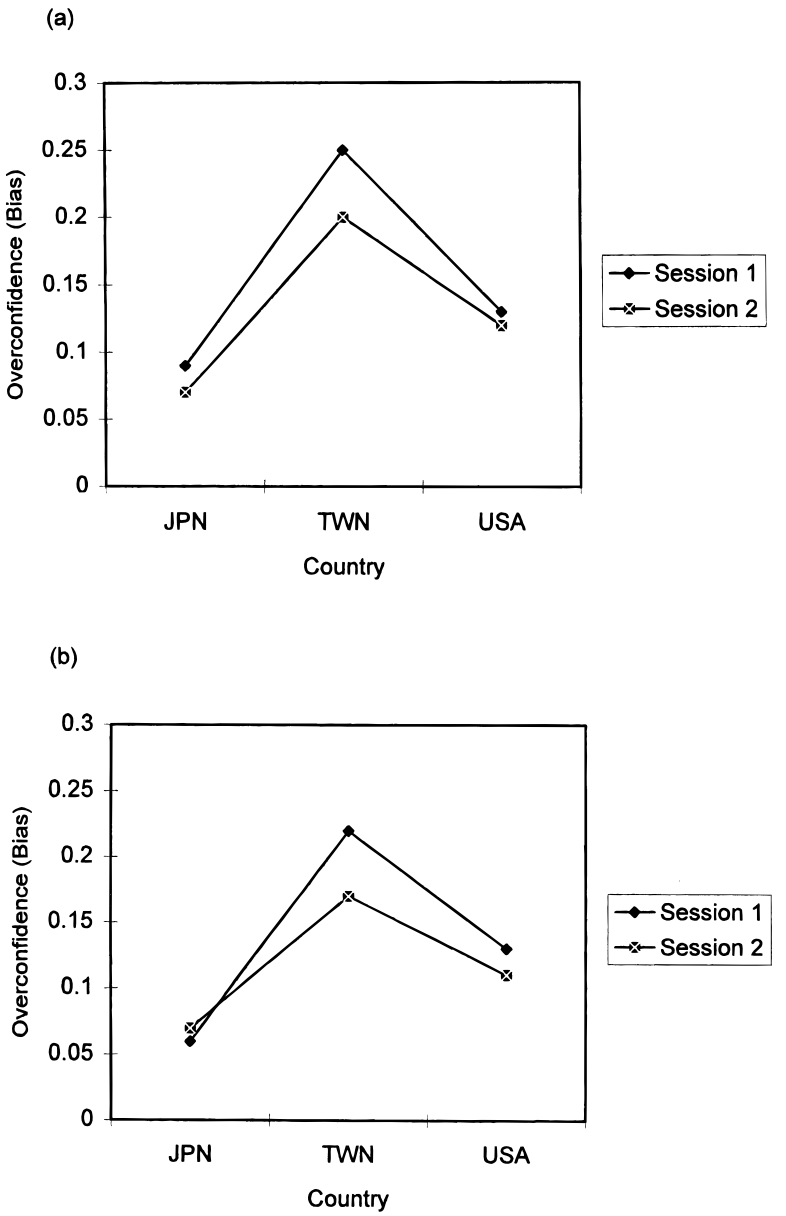

FIG. 2. Mean measures of overconfidence (Bias, internally defined target) by country and session: (a) Study 1 (fixed symptom display), (b) Study 2 (discretionary symptom display). Note: JPN, Japan; TWN, Taiwan; USA, United States.

A second effect that is immediately apparent in Fig. 2 is that the overconfidence demonstrated by the Chinese subjects in Taiwan was substantially greater than that of the Japanese or the American subjects, $t(95) = 6.40$ and 4.38, $p < .001$, for TWN vs JPN in Sessions 1 and 2, respectively; $t(100) = 5.06$ and 3.02, $p < .005$, for TWN vs USA in Sessions 1 and 2, respectively, consistent with what is found in general knowledge studies. Although the comparison was not statistically significant in either instance, in both sessions

the mean level of Japanese overconfidence was lower than that of the Americans, $t(85) = 1.80$ and 1.86, for Sessions 1 and 2, respectively.

There was not an overall interaction between country and session on the bias measure. Nevertheless, whereas for the Japanese and Americans bias was nonsignificantly different across Sessions 1 and 2, for the Chinese in Taiwan, overconfidence was substantially lower in the second session than it was initially, $t(55) = 3.88$, $p < .001$. That is, experience had a marked effect on Chinese but not Japanese or American overconfidence.

Discrimination

As suggested by the earlier discussion of discrimination measures, the discrimination indexes we analyzed were applied to the subjects' 0–100% probability judgments for the externally defined target event A = "This patient has Trebitis." Overall, as depicted in Panel (a) of Fig. 3, from the initial to the final session in Study 1, subjects greatly improved their ability to distinguish instances in which patients had Trebitis from those when they had Philiosis, $F(1, 280) = 23.18$, $p < .001$. There were also substantial country effects, $F(2, 280) = 13.74$, $p < .001$, with Chinese discrimination being consistently the weakest, particularly during the judgment policy formation activities that presumably occurred during Session 1, $t(95) = 4.66$, $p < .001$, and $t(95) = 3.62$, $p < .005$, for TWN vs JPN in Sessions 1 and 2, respectively; $t(100) = 3.24$, $p < .005$, and $t(100) = 1.16$, *ns*, for TWN vs USA in Sessions 1 and 2, respectively. Thus, there are no indications at all of superior Chinese discrimination for meaningfully externally defined target events.

Consistency

With only six dichotomous symptoms, it is unsurprising that our stochastic ecology contained numerous "duplicate" patients with identical symptom profiles. This allowed us to construct for each subject within each session a test–retest reliability r_{tt}, the correlation between the judgments assigned to those duplicates (16 in Session 1 and 17 in Session 2), thus providing insights into the processes by which subjects arrived at their judgments. As illustrated in Panel (a) of Fig. 4, for all subject groups, consistency increased considerably from the first session to the second, $F(1, 280) = 66.63$, $p < .001$ (all tests performed on Fisher-transformed reliabilities). This is, of course, what we should expect, given that in Session 1 subjects were necessarily experimenting, struggling to learn the ecology and develop a judgment procedure that "worked." Observe, however, that there were also substantial country differences, $F(2, 280) = 7.36$, $p = .001$, with a distinctively high degree of consistency on the part of the Japanese being the main driver of this effect, particularly during the learning activities of Session 1, $t(95) = 3.94$, $p < .001$, and $t(95) = 1.87$, $p = .065$, for JPN vs. TWN in Sessions 1 and 2, respectively; $t(85) = 2.52$, $p < .05$, and $t(85) = 1.34$, *ns*, for JPN vs USA in Sessions 1 and 2, respectively.

Table 1 shows the correlations between the measures of consistency (r_{tt}) and

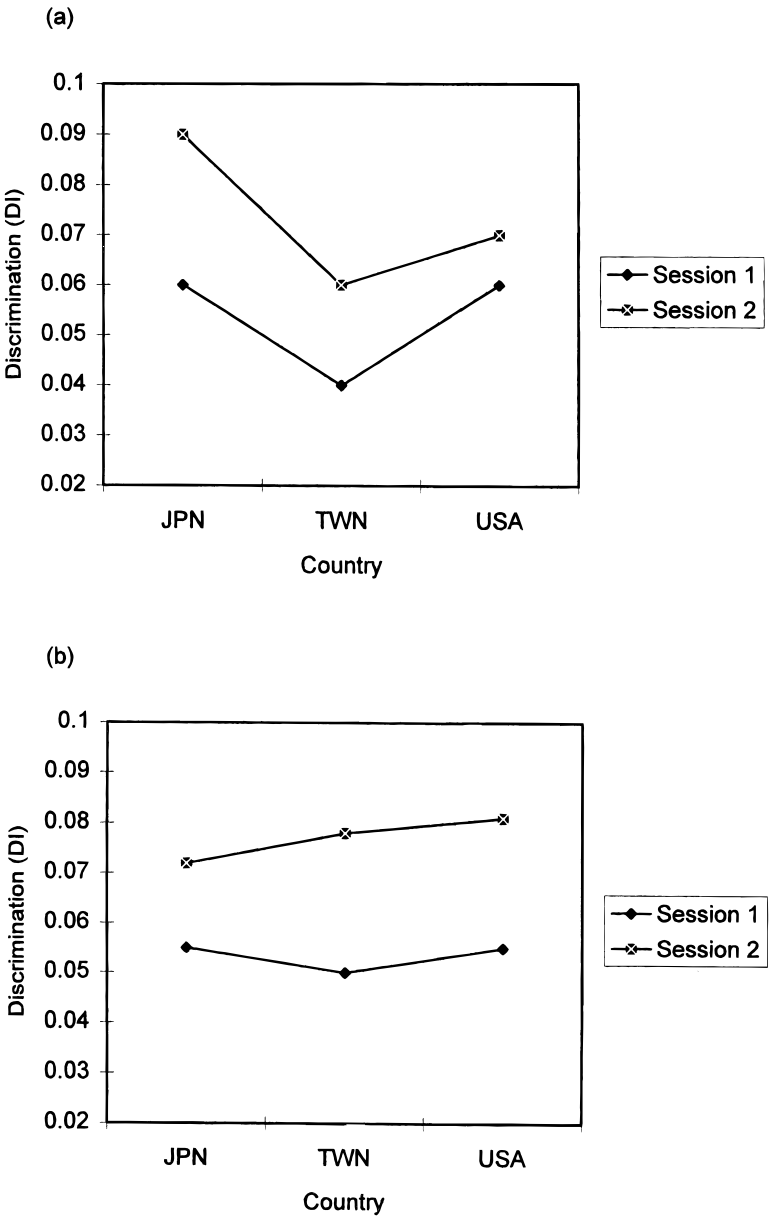

FIG. 3. Mean measures of discrimination (DI, externally defined target) by country and session: (a) Study 1 (fixed symptom display), (b) Study 2 (discretionary symptom display). Note: JPN, Japan; TWN, Taiwan; USA, United States.

overconfidence (bias, with internal target event definition), by country and session. Observe that, in partial agreement with the suggestions by Erev *et al.* (1994), there is some tendency for overconfidence to be associated with inconsistency. Moreover, as indicated above, although the contrasts were not always statistically significant, the Chinese subjects' consistency tended to be weakest among all three groups. Thus, there appears to be modest support for the idea that at least part of the extreme overconfidence exhibited by the

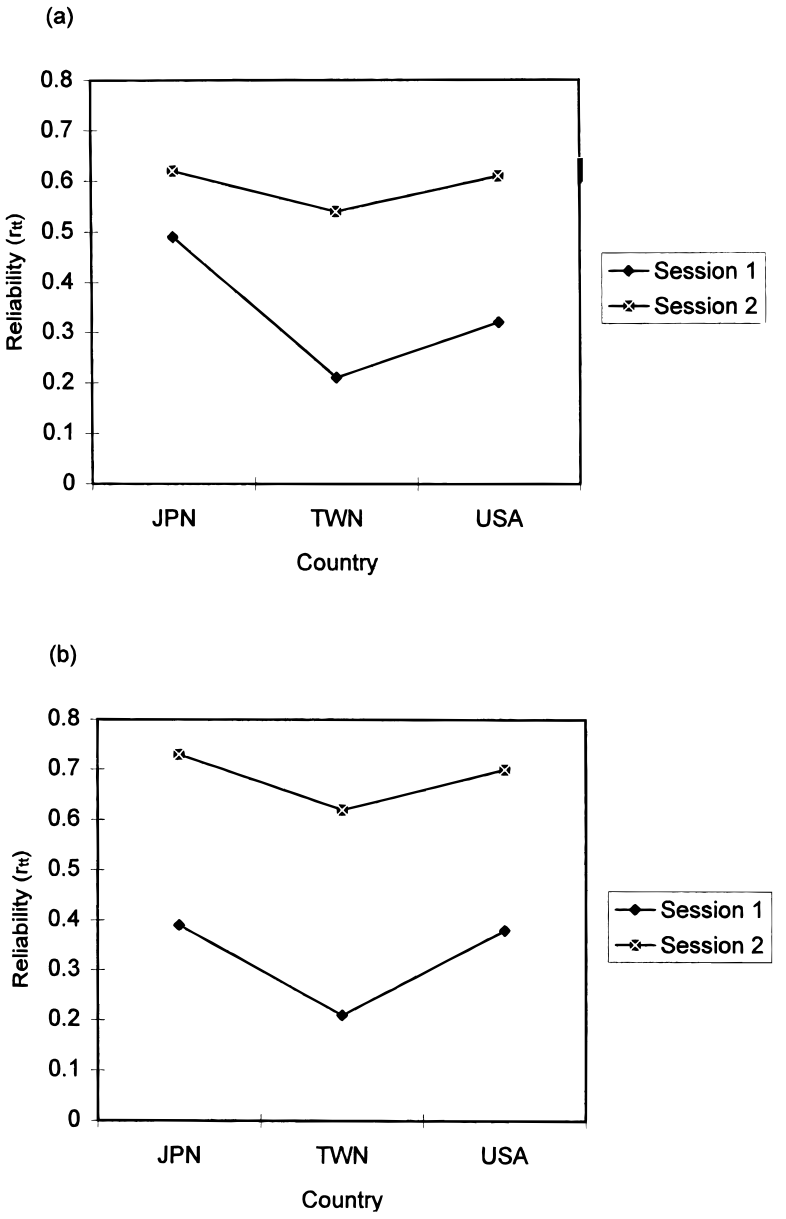

FIG. 4. Mean measures of consistency (r_{tt}) by country and session: (a) Study 1 (fixed symptom display), (b) Study 2 (discretionary symptom display). Note: JPN, Japan; TWN, Taiwan; USA, United States.

Chinese might be mediated by inconsistency. Do inconsistency differences fully account for the observed cross-cultural variations in overconfidence? To test this possibility, we constructed a simple linear model of each subject's judgment policy in each session of Study 1, where all the cues were presented for each patient. We then computed the bias in the "judgments" made by the resulting equations. If cross-cultural variations in overconfidence were completely determined by inconsistency differences, then the biases of these "bootstrapped"

TABLE 1

Correlations between Measures of Consistency (r_{tt}) and Overconfidence (Bias, Internally Defined Target): Study 1 (Fixed Symptom Display)/Study 2 (Discretionary Symptom Display)

	Country		
Session	Japan	Taiwan	United States
1. Development	−.51/−.17	−.25/−.22	−.36/−.36
2. Application	−.55/−.16	−.43/−.60	−.40/−.05

models (cf. Camerer, 1981) should have been essentially the same across all three countries.[3] The biases of the subjects' linear models were much lower than those of the subjects themselves. However, there remained marked country differences in bias, $F(2, 280) = 20.06$, $p < .0001$, in the same basic pattern as before, i.e., with extreme Chinese overconfidence.

Deliberation Time

Figure 5 shows the mean amounts of time subjects took with each patient, from when the patient was first presented to when a probabilistic diagnosis was submitted for that patient. There were large country differences in those deliberation times ($F(2, 280) = 19.69$, $p < .001$) as well as significant session ($F(1, 280) = 100.95$, $p < .001$) and interaction effects ($F(2, 280) = 3.22$, $p < .05$). The session effect is uninteresting, reflecting nothing more than that in the first session subjects were taking the time to develop their judgment policies. The country effect and interaction are important, though. The Japanese subjects were always the most deliberative (although not reliably more than the Chinese in Session 2), but this was especially the case during the learning phase of the study, $t(95) = 3.29$, $p < .001$, and $t(95) = .77$, *ns*, for JPN vs TWN in Sessions 1 and 2, respectively; $t(85) = 5.03$, $p < .001$, and $t(85) = 3.71$, $p < .001$, for JPN vs USA in Sessions 1 and 2, respectively. And the Americans were consistently the least deliberative, $t(100) = 2.07$, $p < .05$, and $t(100) = 3.36$, $p < .001$, for TWN vs USA in Sessions 1 and 2, respectively.

STUDY 2: DISCRETIONARY SYMPTOM DISPLAYS

One aim of Study 2 was to assess the generalizability of the conclusions indicated in Study 1 to a set of conditions that are common in real life. Under those conditions, the person is not automatically presented with a given collection of facts about the case for which a judgment must be rendered. Instead, information sources must be actively queried, as in a clinical interview with a patient. In addition, however, Study 2 was also intended to permit conclusions

[3] We are indebted to Ido Erev for suggesting this analysis.

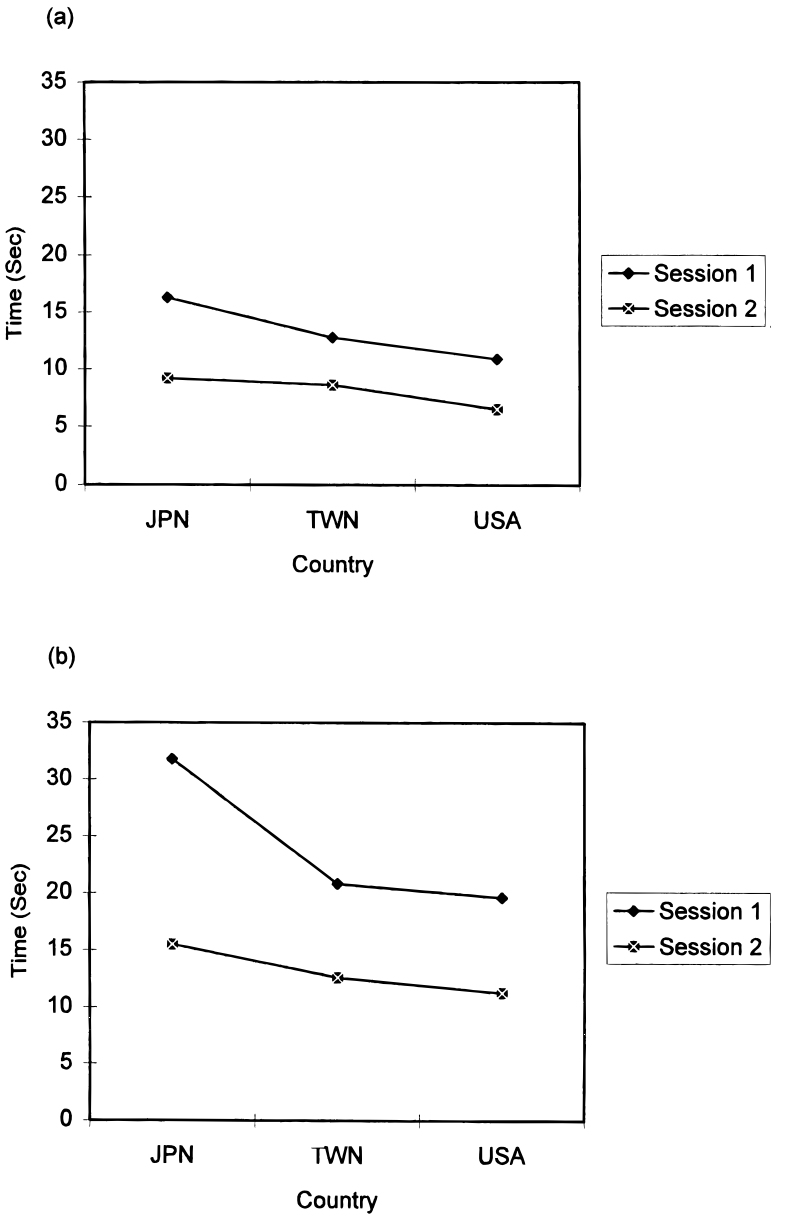

FIG. 5. Mean patient deliberation times (seconds) by country and session: (a) Study 1 (fixed symptom display), (b) Study 2 (discretionary symptom display). Note: JPN, Japan; TWN, Taiwan; USA, United States.

about proposals that the contributors to cross-cultural variations in overconfidence include differences in customs for considering different classes and amounts of information in deliberations. Yates *et al.* (1992), for example, found indications that extreme Chinese overconfidence for general knowledge is due at least in part to a marked disposition against bringing to mind arguments that disagree with their chosen answers to given questions.

Method

The method for Study 2 was the same as that for Study 1 except for the following:

- The numbers of subjects in Taiwan, Japan, and the United States were, respectively, 40, 42, and 42.
- The full six-symptom profile for a given patient was not displayed automatically. Instead, symptoms were presented only upon the subject's request; the subject could ask for as many or as few as desired.
- To maintain fidelity with most naturalistic diagnostic situations, costs were attached to information requests. Specifically, the subject was told that the score used to determine whether the subject received the $10 bonus would be his or her accuracy score adjusted for the amount of information requested, relative to that demanded by the other subjects.
- Whereas in Study 1 the display always presented symptoms in the (unknown-to-the-subject) decreasing validity order described above, they were listed in a single random order in Study 2.
- At the end of the Session 2, the subject was asked a series of postexperimental questions, including requests for estimates of the base rates for the diseases and symptoms and judgments from which the subject's beliefs about symptom diagnosticities could be inferred. For the latter, the subject was asked for the following probability judgments for each symptom: P' (Trebitisl|Symptom Present) and P' (Trebitisl|Symptom Absent), whose difference provides a subjective version of the contingency statistic for that symptom, $\Delta P'$.

Results and Discussion

Panels (b) of Figs. 1–5 as well as the second sets of correlations in Table 1 present the same measures for Study 2 that were used to characterize judgments in Study 1. For the most part, the displays speak for themselves. In the interests of brevity, here we only bring attention to features of the data that implicate key similarities and differences between conclusions for the two types of situations, those in which information displays are fixed vs discretionary.

Overall Accuracy

In comparing Panels (a) and (b) of Fig. 1, the most immediately obvious difference between the two studies concerns overall accuracy in Session 2, after subjects had settled into their judgment routines. Recall that, in Study 1, the overall accuracy of the Chinese subjects was substantially worse than that of the Japanese and the Americans. Those differences in $\overline{PS}$ were sharply reduced in Study 2, $t(80) = 1.84$, *ns*, and $t(82) = 1.88$, *ns*, for TWN vs JPN and TWN vs USA, respectively. This suggests that allowing for discretion in selecting cues is especially helpful for the overall accuracy of the Chinese or, alternatively, that perhaps fixed displays pose special difficulties for judges who adopt Chinese judgment strategies.

Overconfidence

Figure 2 shows a bias pattern that is quite consistent across the two studies. As was the case in Study 1, in Study 2 there was not a statistically significant interaction between country and session, $F(2, 242) = 1.88$, *ns*. Nevertheless, in sharp contrast to the minimal session effects for the Japanese and Americans, bias was substantially reduced in the second session for the Chinese in Taiwan, $t(41) = 2.58$, $p < .01$. That is, once again, there is evidence that experience has an especially strong effect on reducing the overconfidence of the Chinese.

Discrimination

Figure 3 reveals the marked effect that discretion had on the patterns of discrimination effects across countries. Recall that, in Study 1, where the subject was automatically shown all the symptoms for every patient, there were large cultural differences in discrimination, with Chinese discrimination being especially weak. Yet, in Study 2, where the subject had to request the symptoms used to inform his or her diagnoses, all country effects were washed out, $F(2, 242) = .51$, *ns*. As the figure suggests, the apparent basis for this dramatic change was, once again, the positive influence of discretion on the Chinese subjects' judgment quality.

Consistency

As shown Fig. 4, the cross-cultural variations in subjects' judgment consistency were virtually identical in Studies 1 and 2. That figure does, however, illustrate well an incidental across-the-board effect of discretion on subjects' judgments during Session 2, once they had presumably settled into their routines: greater consistency (Bayesian posterior probabilities for higher mean values of r_{tt} in Study 2 vs Study 1 were .980, .924, and .983 for JPN, TWN, and USA, respectively; see Lee, 1989). One plausible potential explanation for this effect is that, with fixed displays, people try to take all available information into account, but the processing demands of such a task are too high for it to be accomplished as reliably as they might wish. An interesting feature of this account is that not all the available symptoms here really were useful in diagnosing the focal diseases, and the subjects did not seem to believe they were either; on average, they requested only three or four of the six symptoms for a given patient. This suggests that, when people are automatically presented with a body of information, they simply assume that it is all pertinent, despite being told (as they were here) that some of it might be useless. Or, perhaps people are simply *incapable* of ignoring a salient item of information even when they think it has no value.

Deliberation Time

It is no surprise that, as shown in Fig. 5, subjects took much longer to make their diagnoses in Study 2 than in Study 1, since they had to choose which

symptoms to examine and then physically select them on the computer. Beyond that, however, the basic cross-cultural patterns in the studies were quite similar though not identical. Most notably, once again the Japanese subjects were significantly more deliberative than the Chinese and the Americans, especially during Session 1, when they were establishing their judgment routines.

Symptom Requests

Table 2 shows the mean numbers of symptoms requested by the subjects. Understandably, there was some tendency for the subjects to ask for fewer cues in Session 2 than in Session 1, even though the session effect was not statistically significant, $F(1, 242) = 2.92$, *ns*. There was, however, a significant country effect, $F(2, 242) = 4.28$, $p < .05$, with no significant interaction. A simple generalization from previous general knowledge findings (Yates *et al.*, 1992) would have predicted that the Chinese subjects would request the fewest symptoms. But that did not happen; the American subjects had that distinction. There were also differences in subject groups' tendencies to request particular symptoms. Yet, there was no discernible pattern to those difference. Hence, we cannot draw defensible conclusions about symptom request appropriateness.

Contingency and Base Rate Beliefs

There were no statistically significant country effects for subjects' judgments about the associations between various symptoms and actual diseases among patients, or for their estimates of symptom base rates. But there were such differences for estimates of the base rate for Trebitis, whose actual value was .60, $F(2, 242) = 12.26$, $p < .001$. The mean estimates for the Japanese, Chinese, and American subjects were, respectively, .58, .67, and .68. The Japanese estimates were clearly far more accurate than those of the other subjects. Oddly enough, however, base rate estimates were almost completely uncorrelated with all the various measures of actual probability judgment accuracy, including bias for externally defined targets. Similar independence has been observed in other studies (e.g., Yates & Estin, 1996) and is reminiscent of Kahneman and Tversky's (1972) proposal that judgments via mechanisms such as the representativeness heuristic reserve no role for a person's assumptions about base rates.

TABLE 2

Mean Numbers of Symptoms Examined for Each Patient, Study 2 (Discretionary Symptoms Display)

	Country		
Session	Japan	Taiwan	United States
1. Development	4.2	4.1	3.7
2. Application	3.9	3.9	3.4

GENERAL DISCUSSION

At the outset of this article, we described a series of key issues the present research was intended to enlighten. Here we briefly recapitulate those questions (although not in the order they were originally posed) and summarize the conclusions the data imply for them. We close with remarks about practical implications.

Almost all published studies on overconfidence have approached the phenomenon using general knowledge questions. This thus motivated the question of the extent to which previous indications of pervasive overconfidence apply to other kinds of probability judgments also, particularly for the kinds of events that bear upon common practical decisions. The simulations in the present studies concerned only one type of alternative scenario, albeit a fairly realistic one that was closely controlled. The results strongly support generalizability. In fact, the overconfidence found here was markedly stronger than what is typically seen in general knowledge studies, and it cannot be attributed to artifacts like biased item selection, which have been shown to account for at least some of the overconfidence observed in general knowledge investigations (e.g., Gigerenzer, Hoffrage, & Kleinbolting, 1991).

Then there are our questions about culture. The first concerned the existence and nature of cross-cultural variations in the overall accuracy of probability judgments. In contrast to what has been reported for general knowledge judgments (Yates *et al.*, 1989), the present data suggest that we should expect substantial differences in judgment accuracy among Japanese, Chinese, and American respondents, provided that all the available cues are brought to the judge's attention. On the other hand, if the judge must actively acquire that information from specified sources, our results suggest that these differences will largely disappear—after Chinese judges have had the opportunity to adapt to their task. Similar conclusions were indicated for cross-cultural variations in discrimination, the ability of a person's judgments to sharply distinguish occasions when a target event is going to occur from those when it is not. That is, large differences should be anticipated when potentially—although not necessarily actually—diagnostic information is routinely put before the judge, with Chinese discrimination weaker than that of Japanese and Americans, but these differences should go away if the judge must seek out the pertinent facts in a given case.

Historically, the main focus of cross-cultural comparisons of probability judgment accuracy has been on overconfidence. Perhaps *the* most important conclusion of the present studies is that cross-cultural variations in overconfidence are not limited to general knowledge; they should be expected in the kinds of judgments that drive common practical decisions too. Overconfidence tends to be especially strong, it seems, in Chinese cultures. And there are indications that it is weakest among the Japanese.

Our data are at least suggestive of several plausible explanations for (or contributors to) the cross-cultural variations in overconfidence that have been found repeatedly. First of all, it is important to recognize what are *not* viable

accounts. Previous work has demonstrated convincingly that the pertinent variations do not rest on affective mechanisms, despite the popularity of affect hypotheses among laypersons as well as scholars (Lee *et al.*, 1995; Yates, Lee, & Shinotsuka, 1996). That is, cultural differences in overconfidence like those examined here cannot be explained in terms of different groups' tendencies to think highly of their personal abilities, perhaps as a means of maintaining high self-esteem. Recent evidence also provides no support for the suggestion that the extreme overconfidence exhibited by some groups, such as the Chinese, reflects nothing more than an extreme response bias (Yates, Lee, & Bush, 1997). The present research adds to the list of proposals that can be ruled out: Given the high degree of control permitted by the procedures used here, there is no reason to suspect that cross-cultural variations in overconfidence are due to differences in previously acquired substantive knowledge about the topics under consideration. The present data also offer no support for the proposition that weak Chinese calibration (e.g., extreme overconfidence) results from an emphasis on achieving good discrimination, leaving little attention to be devoted to attaining good calibration.

Our reliability data provide modest support for one proposed contributor to cross-cultural variations in overconfidence. In agreement with the thesis of Erev *et al.* (1994), that apparent overconfidence can be mediated by inconsistency in articulating underlying "true" judgments, across all subject groups, we found moderate negative correlations between overconfidence and reliability indexes. And, in addition to exhibiting relatively high degrees of overconfidence, our Chinese subjects also tended to be relatively inconsistent in their judgments. This was especially the case early on in our procedures, when overconfidence differences were most pronounced. Of course, the question that immediately comes to mind is an essential one: Why should the Chinese be comparatively inconsistent?

One plausible answer to this question has especially broad implications: Being required to make probability judgments and, perhaps more generally, to engage in the kind of reasoning demanded in such tasks, is relatively uncommon in Chinese cultures. If so, then Chinese inconsistency would be unsurprising; on the spot, Chinese would have to, perhaps through trial and error, develop means of coping with alien requirements that are more familiar to Japanese and Americans. This account also agrees with the relatively greater improvements our Chinese subjects made between the initial and final sessions of our studies. Note that this explanation for judgment inconsistency differs from that suggested by Erev *et al.* (1994). Following the classical test theory model, these authors represent an explicitly articulated probability judgment as a combination of a true score (the person's underlying actual opinion) and an error term. The variance of the error term thus represents the degree of inconsistency in reported judgments. Conceptually, then, inconsistency results from error in the translation between true and articulated opinions. The present familiarity explanation seems more intutively compelling to us than the "errorful translation" proposal; we are unaware of any reason to expect Chinese

culture to make true score translation per se particularly error-prone. But only pointed experimentation will permit a discriminative test of these hypotheses.

What are the important differences in "cognitive customs" (Yates & Lee, 1996) that might provide more positive accounts for cross-cultural variations in overconfidence? One possibility is implicit in Liu's (1986) suggestion that children in Chinese cultures are pointedly taught specific "rules" for approaching various cognitive tasks. Liu's "Rule 6" seems particularly pertinent here (p. 80): "If the purpose is to acquire the knowledge contained in an article, then the best strategy is to memorize the article." Initially, at least, such a strategy might seem quite viable in the tasks used here. That is, the subject might attempt to memorize specific cases (e.g., "Let me keep in mind that Patient 32 had nausea and earache, but no rash . . . and also had Philiosis"), thereby providing the basis for subsequent diagnoses. Suppose that people fail to realize that recall is reconstructive but instead assume erroneously that it entails reading veridical memory traces (e.g., "If I *remember* seeing it, then I really *must* have seen it"). Then, as suggested by a number of authors (e.g., Fischhoff *et al.,* 1977), extreme overconfidence by people who emphasize a memory-intensive approach to tasks like ours (e.g., the Chinese) should follow.

Another Chinese cognitive custom might be implicated as well. Scholarship undertaken from a variety of perspectives (e.g., Munro, 1985; Peng & Nisbett, 1997; Yang, 1986) has demonstrated a Chinese proclivity for holism, a tendency to perceive and conceive of things in terms of wholes rather than parts. Such holism is, incidentally, thought to be especially prominent in Chinese beliefs that the human body is a tightly integrated complete system that should be treated as such in the event of illness (and hence practices such as acupuncture). Thus, Chinese subjects would not have been naturally inclined toward attempting to infer individual symptom validities and then building a subjective diagnostic algorithm relying on those validities. Instead, they might well have been disposed toward the kind of holistic strategy implicit in the kind of memorization approach just described—until its practical limits became apparent. The realization of those limits in a stochastic ecology like that used here should be enhanced by the kind of immediate feedback we provided. And it seems plausible that such a realization would be particularly disconcerting for a person attempting a memorization strategy, fueling the kind of heightened inconsistency observed in our Chinese subjects' responses.

Earlier, we mentioned evidence (Yates *et al.,* 1992) for yet another set of cognitive custom differences that arguably contribute to now-familiar variations in overconfidence. Specifically, that evidence suggests that extreme Chinese overconfidence in general knowledge is supported by the relative rareness with which Chinese culture demands that people generate multiple arguments on both sides of any issue that arises. In this view, overconfidence is the result of the person failing to bring to mind arguments that disagree with the answer to a general knowledge question the person actually selected. On the face of it, the present findings seem to contradict this proposal, in that our Chinese subjects in Study 2 chose to examine as many symptoms as did our American subjects. We resist abandoning the idea so readily, however. The reason is that,

in argument generation, a person must create arguments "from scratch." In contrast, in our Study 2, the display always reminded the subject of the six specific symptoms that were available, placing no demand at all on the subject to think divergently. Indeed, the marked sensitivity of the Chinese to the key distinction between Studies 1 and 2—fixed vs discretionary symptom displays—is consistent with the significance to Chinese cognition of the difference between what is apparent and what must be unveiled. Of course, only additional experiments can determine whether the argument generation hypothesis really should be abandoned.

What is the practical significance of the present findings? At minimum, they imply that in situations where decisions are made using the logic underlying common Western ways of construing decision problems, ignoring the cultures of decision makers is risky; our initial fictional example illustrates what could occur. But depending on the validity of the kinds of fundamental accounts just discussed, the implications might be more extensive. Pollock and Chen (1986) were puzzled and dismayed at their Chinese collaborators' polite indifference to essential decision analytic ideas. More than a decade later, we see a dramatic quickening in the pace of Chinese–Western collaborations of all sorts, including the teaching of Western management techniques in Chinese businesses and business schools. Pollock and Chen's experiences, as well as the present analyses, suggest that such collaborations could be destined for serious difficulties. That is because those efforts might be ignoring fundamental differences in how people in various cultures conceive of decision making. (Trompenaars, 1993, describes numerous examples of how other varieties of cultural differences create formidable and sometimes fatal barriers to effective collaborations.) Thus, Chinese decision makers might judge probabilities in their distinctive way because the customary Chinese construal of a decision problem does not conform to the metaphor underlying prevailing Western decision schemas (e.g., one representable by a decision tree). Zhang (1992), for one, has proposed that this is, in fact, the case, that characteristically Chinese decision strategies emphasize a quite different logic, that of historical precedence.

Another incidental aspect of the present results might also have noteworthy practical import in a different quarter. Recall that our Japanese subjects were markedly slower than the other subjects, especially the American subjects in the initial sessions of the studies, when subjects presumably were learning the ecology and formulating their judgment strategies. There is reason to believe that this distinctive Japanese deliberativeness is not peculiar to the present procedures. Instead, it might well reflect a more general Japanese cultural norm for thoroughness, one that is explicitly encouraged in Japanese educational practices. Consider, for instance, Stevenson and Stigler's (1992) observational studies of Japanese elementary school mathematics classrooms, which reveal a pronounced "Japanese emphasis on reflection" (p. 195). The contrast with American norms is especially telling in the following report by Stevenson and Stigler (p. 9): "American elementary school students, watching a videotape of a Japanese mathematics lesson, inevitably react to the pace: They perceive

unbearable slowness." Now imagine a situation in which Japanese and Americans are attempting to make judgments and decisions collaboratively. If, as the data suggest, there really are substantially different Japanese and American norms for thoroughness as opposed to speed, the complications—including mutual disdain—are apparent. Which emphasis, thoroughness or speed, is objectively better? That is difficult to say. In the present studies, Japanese-style deliberativeness was no more effective in yielding accurate judgments than American-style cursoriness. But we doubt that this will always be true. Our hunch is that relative effectiveness depends on as yet unspecified characteristics of the circumstances.

REFERENCES

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review,* **78**(1), 1–3.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology,* **1,** 185–216.

Camerer, C. (1981). General conditions for the success of bootstrapping models. *Organizational Behavior and Human Performance,* **27,** 411–422.

Clemen, R. T. (1991). *Making hard decisions: An introduction to decision analysis.* Boston: PWS–Kent.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review,* **101,** 519–527.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance,* **3,** 552–564.

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review,* **98,** 506–528.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology,* **3,** 430–454.

Lee, J.-W., Yates, J. F., Shinotsuka, H., Singh, R., Onglatco, M. L. U., Yen, N. S., Gupta, M., & Bhatnagar, D. (1995). Cross-national differences in overconfidence. *Asian Journal of Psychology,* **1,** 63–69.

Lee, P. M. (1989). *Bayesian statistics: An introduction.* New York: Halsted Press.

Levi, K. R. (1986). *Numerical likelihood estimates from physicians and linear models.* Doctoral dissertation, University of Michigan, Ann Arbor.

Liu, I.-M. (1986). Chinese cognition. In M. H. Bond (Ed.), *The psychology of the Chinese people* (pp. 73–105). Hong Kong: Oxford University Press.

Munro, D. J. (Ed.). (1985). *Individualism and holism: Studies in Confucian and Taoist values.* Ann Arbor: University of Michigan Center for Chinese Studies.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology,* **12,** 595–600.

Peng, K., & Nisbett, R. E. (1997). *Cultural influences on perception of covariation.* Unpublished manuscript under editorial review, Department of Psychology, University of Michigan, Ann Arbor.

Pollock, S. M., & Chen, K. (1986). Strive to conquer the Black Stink: Decision analysis in the People's Republic of China. *Interfaces,* **16**(2), 31–37.

Stevenson, H. W., & Stigler, J. W. (1992). *The learning gap: Why our schools are failing and what we can learn from Japanese and Chinese education.* New York: Summit.

Tape, T. G., Heckerling, P. S., Ornato, J. P., & Wigton, R. S. (1991). Use of clinical judgment

analysis to explain regional variations in physicians' accuracies in diagnosing pneumonia. *Medical Decision Making,* **11,** 189–195.

Trompenaars, F. (1993). *Riding the waves of culture.* London: The Economist Books.

Whitcomb, K. M., Onkal, D., Curley, S. P., & Benson, P. G. (1995). Probability judgment accuracy for general knowledge: Cross-national differences and assessment methods. *Journal of Behavioral Decision Making,* **8,** 51–67.

Wright, G., & Ayton, P. (1986). Subjective confidence in forecasts: A response to Fischhoff and MacGregor. *Journal of Forecasting,* **5,** 117–123.

Wright, G. N., Phillips, L. D., Whalley, P. C., Choo, G. T., Ng, K. O., Tan, I., & Wisudha, A. (1978). Cultural differences in probabilistic thinking. *Journal of Cross-Cultural Psychology,* **9,** 285–299.

Wright, G., & Wisudha, A. (1982). Distribution of probability assessments for almanac and future event questions. *Scandinavian Journal of Psychology,* **23,** 219–224.

Yang, K.-S. (1986). Chinese personality and its change. In M. H. Bond (Ed.), *The psychology of the Chinese people* (pp. 106–170). Hong Kong: Oxford University Press.

Yaniv, I., Yates, J. F., & Smith, J. E. K. (1991). Measures of discrimination skill in probabilistic judgment. *Psychological Bulletin,* **110,** 611–617.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance,* **30,** 132–156.

Yates, J. F. (1990). *Judgment and decision making.* Englewood Cliffs, NJ: Prentice Hall.

Yates, J. F. (1994). Subjective probability accuracy analysis. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 381–410). Chichester, England: Wiley.

Yates, J. F., & Estin, P. A. (1996, November). *Training good judgment.* Paper presented at the Annual Meeting of the Society for Judgment and Decision Making, Chicago.

Yates, J. F., & Lee, J.-W. (1996). Chinese decision making. In M. H. Bond (Ed.), *Handbook of Chinese psychology* (pp. 338–351). Hong Kong: Oxford University Press.

Yates, J. F., Lee, J.-W., & Bush, J. G. (1997). General knowledge overconfidence: Cross-national variations, response style, and "reality." *Organizational Behavior and Human Decision Processes,* **70,** 87–94.

Yates, J. F., Lee, J.-W., & Shinotsuka, H. (1992, November). *Cross-national variation in probability judgment.* Paper presented at the Annual Meeting of the Psychonomic Society, St. Louis.

Yates, J. F., Lee, J.-W., Levi, K. R., & Curley, S. P. (1990). Measuring and analyzing probability judgment accuracy in medicine. *Philippine Journal of Internal Medicine,* **28** (Suppl. 1), 21–32.

Yates, J. F., Lee, J.-W., & Shinotsuka, H. (1996). Beliefs about overconfidence, including its cross-national variation. *Organizational Behavior and Human Decision Processes,* **65,** 138–147.

Yates, J. F., Zhu, Y., Ronis, D. L., Wang, D.-F., Shinotsuka, H., & Toda, M. (1989). Probability judgment accuracy: China, Japan, and the United States. *Organizational Behavior and Human Decision Processes,* **43,** 145–171.

Zhang, B. (1992). *Cultural conditionality in decision making: A prospect of probabilistic thinking.* Unpublished doctoral dissertation, Department of Information Systems, London School of Economics and Political Science, University of London, London.

Received: July 4, 1997