Over/Under Confidence: Effects of Culture and Number of Options

Bruce D. Burns (bruce.burns@sydney.edu.au)
School of Psychology
University of Sydney, NSW 2006 Australia

Luming Luo (lluo1019@uni.sydney.edu.au)
School of Psychology
University of Sydney, NSW 2006 Australia

Abstract

Greater over-confidence in answers to multiple choice general knowledge questions has been found for people in East Asian countries compared to English-speaking countries. A drawback of this research is difficulty in establishing the equivalence of samples across countries, so we compared students at the same university whose first language was either East Asian, English, or Other. Our earlier research using Chinese speaking students suggested under-confidence; however we had presented questions with four response options rather than two. Therefore here we also manipulated the number of response options. We found that the East Asian group consistently performed worse at a given level of confidence than the other groups, but that they displayed under rather than over-confidence for 4-option items. Thus our results were consistent with findings of greater confidence for people with East Asian roots, but whether this manifests as over-confidence, under-confidence, or better calibration could depend on the question's structure.

Keywords: Confidence, decision making, culture.

Introduction

In today's progressively globalised world, understanding cultural differences in decision-making is a valuable and increasingly relevant area of research. Thus evidence of differences in probabilistic thinking using cross-national samples is important. For example, studies have predominantly found similarities in degrees of over-confidence between English speaking countries, but greater over-confidence in East Asian samples. However methodological issues with this research have been raised and this paper will attempt to address some of these.

Cultural Differences in Probabilistic Thinking

One of the most intriguing findings in cross-cultural research today is the illustration of the existence of cultural differences in cognition and reasoning styles (Ji, Zhang & Nisbett, 2004). Notably, it has been suggested that East Asians reason in a holistic and relational manner, whereas Westerners tend to reason in an analytic fashion (Nisbett, 2003; Nisbett, Peng, Choi & Norenzayan, 2001). It is perhaps unsurprising then, that previous research has also demonstrated qualitative cultural differences in the way people perceive uncertainty and utilise probabilistic information.

The most robust findings of cultural differences in probabilistic thinking are found in the calibration literature (Yates, Lee, Shinotsuka, Patalano & Sieck, 1998). That is, East Asians are generally found to be more overconfident in their probability estimates. This suggests that decision-making under uncertainty may be influenced by cultural factors. Most of this research has examined calibration through the general knowledge format, which suggests that East Asians tend to engage in a more dichotomous style of probabilistic thinking, which would imply a tendency towards responding to uncertainty in either an extreme or conservative fashion. However, there are methodological issues in the general knowledge format with regards to sampling and the number of alternatives provided, the limitations and implications of which will be discussed.

Cultural Calibration of Probabilistic Information

The accuracy of probability judgments is vital as it underpins the quality of decisions (Yates, Zhu, Ronis, Wang, Shinotsuka & Toda, 1989). Lichtenstein and Fischhoff (1977) underscored 'calibration' as a facet of probability assessment accuracy, and subsequently, most research has focused on this aspect of probability judgment. Calibration refers to the extent to which probability judgments regarding the occurrence of certain events are congruent with the relative frequencies of their actual occurrence (Lichtenstein & Fischhoff, 1981). In other words, it measures the degree of concordance between subjective confidence and objective accuracy. Well-calibrated probability assessments are valuable as they facilitate the accurate interpretation of the judgment being communicated, which allows for more reliable decision-making.

Assessing Calibration. General knowledge questions are the most widely used tool for assessing calibration. They generally require the participant to first select an answer from one or more alternatives and then qualify this through indicating a subjective degree of certainty that their choice is correct. They are considered an appropriate method for assessing calibration as they provide an objective measure, where the performance criterion can be clearly defined (Keren, 1991).

The calibration literature reveals two major findings: a susceptibility to give unrealistically high estimates, particularly for general knowledge questions, which is interpreted in the literature as overconfidence (e.g., Fischhoff, Slovic, & Lichtenstein, 1977; Lichtenstein &

Fischhoff, 1977); and an effect for the difficulty of the task on degree of confidence, whereby a more difficult task usually elicits a higher degree of overconfidence (e.g., Arkes, Christensen, Lai & Blumer, 1987; Keren, 1991). The latter phenomenon has been labeled the 'hard-easy effect' (Gigerenzer, Hoffrage & Kleinbolting, 1991). Both of these findings are fairly robust and studies which have attempted to ameliorate this through varying response format, instructions given and offering rewards, have still found the same effects (e.g., Lichtenstein & Fischhoff, 1977; Lichtenstein & Fischhoff, 1981).

Main Findings in Cross-Cultural Comparisons of Calibrations. The most consistent finding in the cross-cultural calibration literature is that East Asians demonstrate a more marked overconfidence in general knowledge tasks, as reflected in their calibration curves. Phillips and Wright (1977) developed the Probability Assessment Questionnaire (PAQ), designed to give an assessment of whether probability estimates are well calibrated. The PAQ is in the format of a two-alternative general knowledge questionnaire, where participants select their answer and then provide a confidence estimate on a 50-100% scale. They found that in the PAQ, British participants produced less extreme probabilities and were better calibrated than the Asian participants, who demonstrated a greater degree of overconfidence in their judgments. Moreover, these differences could not be predicted through the availability of probabilistic expressions in the different languages. These findings were also corroborated by those of Wright and Phillips (1980), and Wright and Phillips (1978), who found that although their British, Hong Kong, Malaysian and Indonesian sample all demonstrated overconfidence in the PAQ, the British sample were the most accurately calibrated.

The most dominant explanation for these differences is the claim that performance on calibration tasks reflects a tendency by East Asians to adopt a 'black-and-white' style of thinking, whereby the world is viewed dichotomously in terms of certainty or total uncertainty, whilst the Westerners are more predisposed to engage in 'probabilistic thinking' where the world is viewed in terms of degrees of probability (Wright, 1981). It is this difference which is assumed to account for a tendency of Asian subjects to use 100% assessments overconfidently, which could be labeled as the 'certainty illusion' (Wright, 1981).

Extending Cross-National Studies

Much of the previous literature exploring cultural differences in cognitive processing has used largely homogenous cross-national samples (Benet-Martinez, Leu, Lee & Morris, 2002). However, this is subject to difficulties in controlling for important variables such as sample characteristics, task equivalence, and experimenter attributes. A novel way of addressing these issues and controlling for these variables is the use of a bicultural sample.

Bicultural individuals are of particular interest because they complement cross-cultural comparisons in helping to isolate the causal role of culture, whilst allowing for greater internal validity through permitting greater equivalence of groups (Hong, Morris, Chiu & Benet-Martinez, 2000). This defines bicultural individuals as those who have internalised two cultures to the extent that both have influences on their thoughts, feelings and behaviour (Hong et al., 2000). One way of defining biculturalism has been in terms of bilingualism, though this is not unproblematic (Portes & Rumbaut, 2001).

Methodological Issues: Does the Number of Options Matter? The current literature using general knowledge questions as a measure for calibration has a possible limitation in that the majority of general knowledge tasks tend to use a dichotomous response option with a half-scale, where participants must indicate their level of certainty on a 50-100% scale. Few studies have examined confidence judgments on a full-range scale with multiple response options, and in fact, all cross-cultural studies testing general knowledge appear to have used a two-alternative format. The number of alternatives made available to the assessor is an important consideration as it defines the type of probability scale used and the relative significance of each chance level (Keren, 1991). Although the overconfidence effect has been found to remain even if the full-range response scale has been provided (Gigerenzer, Hoffrage, & Kleinbolting, 1991), this has still only been established in a dichotomous response-option format. It has been suggested that probability assessments are sensitive to the number of alternatives provided (Keren, 1991). Pallier, et al (2002) used a 5-response option scenario for a general knowledge task and still found overconfidence in their sample. However Luo (2011) gave Chinese-Australian bicultural university students general knowledge question with four response options and found under-confidence rather than over-confidence in their confidence judgments.

The Present Study. This study had two aims: (i) to test for the existence of overconfidence in a bi-cultural East Asian sample, and (ii) to test the effects of varying the number of alternatives a question presents to participants.

According to the Australian Bureau of Statistics, in 2011 27% of Australians were born overseas and a further 20% had at least one overseas-born parent. Amongst foreign-born Australians, China is the third-most common place of birth. Accordingly, Australian university samples contain a large number of students whose first language is East Asian. However in order to gain admission they must meet the same criteria as other students in terms of academic excellence and English-language competence. Thus they are reasonably matched to their fellow students whose first language is English. Therefore they provide an opportunity to test the effects of culture on over-confidence if we make the assumption that students with an East Asian first-

language will at least to some extent be culturally East Asian, even after living in an English speaking country.

Luo (2011) found that under-confidence characterized a Chinese bicultural sample, but as all her participants responded to 4-option general knowledge questions it was impossible to ascertain whether this inconsistency with previous research was due to their biculturalism or the number of response options. So in this study we systematically varied the number of options to be either 2 or 4. If there are cultural effects on overconfidence, then they should be robust across factors such as number of options.

Method

Participants

Participants were first-year psychology students at the University of Sydney, Australia. A total of 793 completed the experiment, but only the 774 who attempted all 16 general knowledge items were included in analyses in order to avoid introducing any biases due to differences in questions responded to. The average age of the sample was 19.53 years and 65.5% were female. In response to a question about their first language, 445 (57.5%) indicated English and 144 (18.6%) indicated an East Asian language (114 Chinese, 6 Japanese, 25 Korean). A further 185 indicated some other first language, of which the most common were Arabic (35) and Vietnamese (31). Exactly how to define our Asian sample was unclear given that the definition has varied between studies. The most common claims seem to be about people from East Asian counties, so we decided a priori to create the groups we did.

Materials and Procedure

We adopted a modified version of the general knowledge task used in Willaby's (2010) study. We used this particular set of questions because we had access to data showing them to be difficult (mean number of questions correct was 58.14%) and the easy-hard effect predicts that hard questions augment overconfidence effects. Previous studies have provided participants with two alternatives, and as a consequence, the calibration curves have been measured on a 50-100% scale. We were interested in exploring whether the same pattern of overconfidence extended below 50%, and whether the number of response options had an effect. Therefore, half the questions given participants had four options, such that the calibration curve was measured from 25-100%. An example of a 4-option item is:

The Ring of Fire is located around what ocean?
•Atlantic •Pacific •Southern •Indian
Percent probability that chosen answer is correct _____%

The 2-option version of this item presented the correct response (Pacific) and one other randomly selected response option (in this case, Atlantic).

Participants answered 16 general knowledge questions. The 16 items were always presented in the same order, but eight of the items were presented with two alternative answers and eight were presented with four alternative answers. Whether the first or second set of eight questions were 4-option or 2-option items was varied across participants. Thus across participants all questions were presented about equally often with either two or four possible answers. In between answering each item participants made an estimate of a quantity which formed part of a different experiment. The experiment was completed on-line and took under 15 minutes to complete. Participants were urged not to look up the answers to questions, and there is no evidence that they did so.

Results

Two types of calibration have been defined in the literature, calibration-in-the-small and calibration-in-the-large (Yates et al., 1989). We will present the data in both ways.

Calibration-in-the-large

Calibration-in-the-large is a single index measure of judgment accuracy. It refers to the extent, over all occasions, to which the average assigned probability judgment matches the proportion of times that the target event actually occurs (Yates et al., 1989). In relation to general knowledge questions, calibration-in-the-large serves as a better measure of overconfidence than calibration-in-the-small (Yates et al., 1989). It is often calculated using the Bias statistic, which is the difference between the average confidence that the chosen answer is correct and the proportion of correct answers (Yates, 2010).

The Bias statistic represents the degree to which the participants are calibrated overall and is calculated for each participant by subtracting the percentage of questions correct from his or her mean probability estimate across all questions. Rather than calculate just the difference, for greater informativeness we present in Table 1 both the mean percent corrects and the mean probability (confidence) estimates. Given that only participants who attempted every item are included in the analysis, the difference between these two means would be the mean Bias statistic. A positive difference in favour of confidence is generally taken to indicate over-confidence, whilst a negative difference indicates under-confidence. Perfect calibration would be demonstrated by a zero difference.

Table 1 breaks participants into three groups based on what they reported to be their first language: English, an East Asian language, or Other language. A 2x2x3 Mixed-design MANOVA was run with factors for number of item response options (2 vs. 4), bias (%correct vs. %confidence) and first language (English, East Asia, or Other). There was a main effect of option, $F(1,771) = 186.78$, $p < .001$, indicating that participants were both more confident and more likely to be correct when an item was presented with 2 rather than 4 options, which effectively makes an item easier. There was a main effect for bias, $F(1,771) = 43.53$, $p < .001$, demonstrating that overall confidence was lower than the estimated proportion correct, that is, our

participants displayed under-confidence overall. There was a main effect of language, $F(2,771) = 31.37$, $p < .001$, as well as a two-way interaction between language and bias, $F(2,771) = 12.67$, $p < .001$. There was also a two-way interaction of options by bias, $F(2,771) = 15.73$, $p < .001$, but no language by options interaction, $F(2,771) = 0.54$, $p = .582$. The three-way interaction was not significant, $F(2,771) = 0.53$, $p = .590$.

Table 1: Mean percentages correct and percents confidence (SDs in parentheses) for 4-option and 2-option items, for participants whose first language was English, an East Asian language, or Other language.

		4-option items		2-option items	
		Correct	Confidence	Correct	Confidence
First language	English (n=445)	64.55 (24.62)	55.20 (20.55)	76.18 (19.91)	71.90 (14.58)
	East Asian (n=144)	52.00 (26.60)	50.23 (18.70)	64.67 (22.58)	66.25 (13.19)
	Other (n=185)	58.38 (22.55)	52.47 (19.51)	73.38 (20.42)	70.66 (13.89)
	Total (n=774)	60.76 (24.91)	53.65 (20.04)	73.37 (20.97)	70.54 (14.31)

To examine the effect of language, the same analysis was conducted including only the English and East Asian groups. The same pattern of significant results was found, in particular there was an interaction between language and bias, $F(1.587) = 24.48$, $p < .001$. The same analysis with East Asian and Other groups also showed the same pattern of results, again with a significant language by bias interaction, $F(1,327) = 6.63$, $p = .010$. This suggests that the East Asian effect is not simply a bilingual effect.

Calibration-in-the-small

Calibration-in-the-small refers to the degree to which the proportion of correct responses for each subjective confidence category correspond to the mean confidence level represented by each category; this allows for a discrepancy between confidence and accuracy to be calculated for each category (Yates et al., 1989). For example, calibration-in-the small would compare, say, for all cases where the person assigned 60% certainty, how close the proportion of correct answers were to 60%. An identity line can be drawn for which confidence and correctness are in perfect synchrony. In the context of general knowledge questions, overconfidence is evident when the calibration curve sits to the right of the identity line (see Figure 1), whereas under-confidence is evident when it sits to the left.

Although calibration-in-the-large is considered a better measure, previous studies examining cultural differences have predominantly used calibration-in-the-small to demonstrate that East Asian participants showed more marked overconfidence. Calibration-in-the-small can be used to provide an indication of the consistency of under or overconfidence effects. Therefore, in this study we also calculated the calibration curves representing calibration-in-the-small in order to create a point of comparison with other studies. To simplify the graphs only the English and East Asian groups will be shown.

The calibration curve was formed through calculating the average proportion across all questions that a particular category was selected and led to a correct response. For 4-option items nine categories were created for each question: 25%, 25-29%, 30-39%, 40-49%, 50-59%, 60-69%, 70-79%, 80-89%, 90-99% and 100%. The extreme responses (25% and 100%) were represented as their own categories because they may have special meaning for participants.

Figure 1 is consistent with the finding from the calibration-in-the-large measure that the East Asian group tended to perform more poorly for a given level of confidence, but it augments this finding by showing that this is consistent for every level of confidence. However both groups tended to display over-confidence when absolute confidence levels were high but under-confidence when absolute confidence was low. The reason why calibration in-the-small suggests overall under-confidence for the English-language group and East Asian language group was because most of their confidence judgements were low, as is shown in Figure 2.

The calibration curves were calculated for 2-option items using seven categories: 50%, 50-59%, 60-69%, 70-79%, 80-89%, 90-99% and 100%. Again extreme values were their own categories.

Figure 3 demonstrates again that for each confidence judgement category the East Asian group tended to perform more poorly than the English group. However, as the calibration-in-the-large data suggests, this leads the East Asian group towards better calibration rather than over-confidence. As for 2-option items, under-confidence appears to be more likely to be displayed at higher absolute confidence levels than lower ones. The calibration in-the-small finding of overall under-confidence is driven by the predominance of lower confidence judgements shown in Figure 4.

Discussion

Overall our results are consistent with the general finding that, to the extent that first-language reflects culture, an East Asian group will have poorer performance for each given level of performance than a culturally Western group. This finding is an important extension of previous findings because by using a bicultural East Asian group our groups

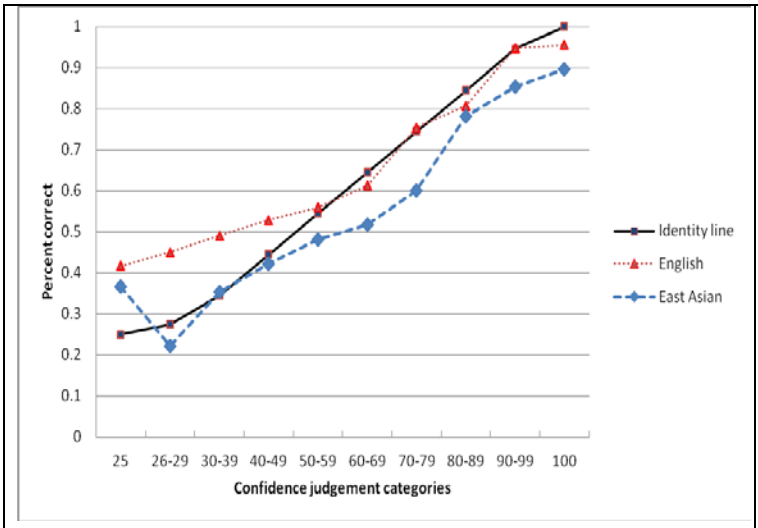

Figure 1: Calibration curves for 4-option items for monolingual and East Asian language speakers. The identity line represents perfect calibration between confidence and correctness.

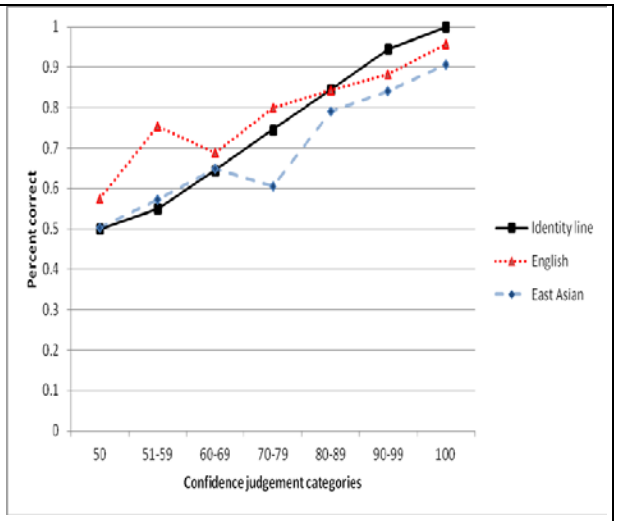

Figure 3: Calibration curves for 2-option items for monolingual and East Asian language speakers. The identity line represents perfect calibration between confidence and correctness.

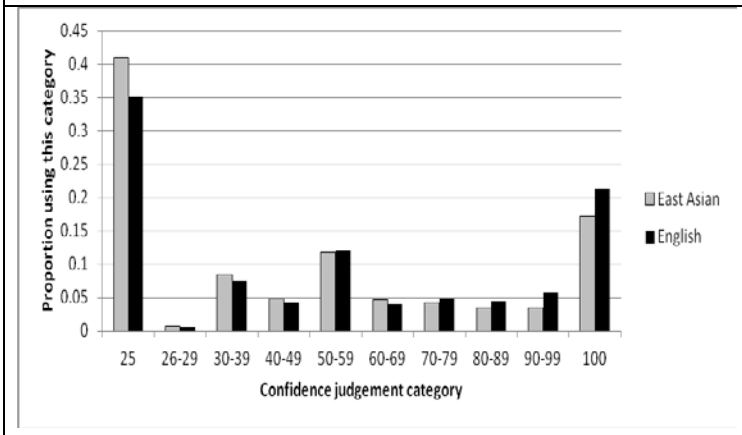

Figure 2: Proportionate use of each confidence category for 4-option items for English and East Asian groups.

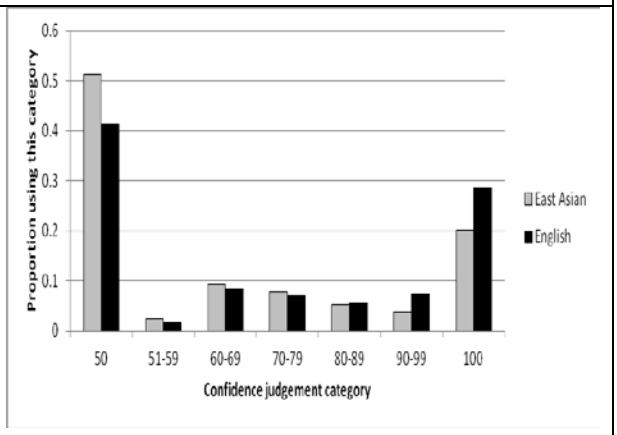

Figure 4: Proportionate use of each confidence category for 2-option items for English and East Asian.

were better matched than in cross-national studies. All our participants had the same materials, similar high levels of English language skills, and the intelligence to be admitted to the same major Australian university. They had all lived in the same country and despite unmeasured variation in how long the East Asian participants had lived in Australia; we still found cultural effects on confidence calibration. The finding that the East Asian group was differentiated from a group with some other non-English first language argues against our findings being due simply to bilingualism or being part of an immigrant group.

The effect of number of options is also an important finding because it suggests that despite the difference we found for the East Asian group, it may not be accurate to characterize this group as having greater overconfidence. Our results suggest that whether the confidence difference manifests itself as over-confidence, under-confidence or better calibration depended on the nature of the task and where you looked on the calibration curve. The general similarity of the groups displayed in the frequency data shown in Figures 2 and 4 also tends to argue against the East Asian sample being more prone to black-white thinking. Thus, our results could contribute to trying to determine the reasons for the cross-cultural differences.

Our overall finding of under-confidence is surprising given that overconfidence is considered as a fairly robust finding in other studies. The findings from this study do not appear to support those found in the calibration literature: a pervasive susceptibility to give unrealistically high estimates, or the 'easy-hard effect', which would predict,

given the demonstrated difficulty of this set of questions, greater evidence of overconfidence at the low ends of the calibration curves. However, few studies have examined confidence judgments on a full-range scale with multiple response options. Pallier et al. (2002) used a difficult general knowledge task which provided five alternatives, with a full scale (20-100%) and also found overconfidence in their sample. Nevertheless, the findings of this study may suggest that when scales are extended through an increase in the number of response options, overconfidence is less likely to occur.

Limitations and Future Directions

The extent to which first language reflects culture for immigrant groups could be challenged, especially because we did not collect any information about culture or how long participants had lived in Australia. We think it a reasonable assumption that first language and culture are correlated, but future research should collect measures of acculturalisation and would predict that such measures would mediate any East Asian confidence effects. Our data also does not allow us to rule out that this effect could be due to language rather than culture.

Why we found evidence of under-confidence in native English-speakers whereas studies in the US and UK have indicated overconfidence is hard to say. We assumed that Australian could be bracketed together with other native English speaking nations, but perhaps this is not the case. It could alternatively be due to something about our items, the online presentation, or the university. If our methodology was responsible for the discrepancy then it adds to the point made by the 2-option/4-options effect, that blanket assertions of over-confidence may fail due to seemingly unimportant variations in how the task is presented. Further examination of such methodological factors may yield clues for understanding miscalibrations of confidence.

References

Arkes, H, R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes, 39*, 133-144.

Benet-Martinez, V., Leu, J., Lee, F., & Morris, M. W. (2002). Negotiating bicuturalism: Cultural frame switching in biculturals with propositional versus compatible cultural identities. *Journal of Cross-Cultural Psychology, 33*, 492-516.

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., & Kleinbolting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review, 98*, 506-528.

Hong, Y.-Y., Morris, M. W., Chiu, C.-Y., & Benet-Martinez, V. (2000). A dynamic constructivist approach to culture and cognition. *American Psychologist, 55*, 709-720.

Lichtenstein, S. B., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20*, 159-183.

Lichtenstein, S. B., & Fischhoff, B. (1981). *The effects of gender and instructions on calibration. decision research*. Technical Report, PTR-1092-81-7, Euguene, OR.

Luo, L. (2011). *Probabilistic thinking in Australian-Chinese biculturals.* Unpublished honours thesis, the University of Sydney.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77*, 217-273.

Ji, L.-Y., Zhang, Z., & Nisbett , R. E. (2004). Is it culture or is it language? Examination of language effects in cross-cultural research on categorisation. *Journal of Personality and Social Psychology, 87*, 57-65.

Nisbett, R. E. (2003). *The geography of thought: How Asians and Westerners think differently…and why*. New York, NY: The Free Press.

Nisbett, R. E., Choi, I., Peng, K., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review, 108*, 291-310.

Pallier, G., Wilkinson, R., Danthiir, V., Kleitman, S., Knezevic, G., Stankov, L., & Roberts, R. D. (2002). The role of individual differences in accuracy of confidence judgments. *The Journal of General Psychology, 129*, 257-299.

Phillips, L. D., & Wright, G. N. (1977). Cultural differences in viewing uncertainty and assessing probabilities. In: H. Jungermann and G. de Zeeuw (eds.), *Decision making and changes in human affairs.* Amsterdam: D. Reidel.

Portes, A., & Rumbaut, R. G. (2001). *Legacies: The story of the immigrant second generation*. Berkley: University of California Press.

Willaby, H. (2010). *Luck feelings, luck beliefs, and decision making.* Unpublished doctoral dissertation, the University of Sydney.

Wright, G. N. (1981). Cultural and task influences on decision making under uncertainty. *Current Anthropology, 22*, 290-291.

Wright, G. N. & Phillips, L. D. (1980). Cultural variation in probabilistic thinking: Alternative ways of dealing with uncertainty. *International Journal of Psychology, 15*, 239-257.

Yates, J. F. (2010). Culture and probability judgment. *Social and Personality Psychology Compass, 4, Supp*, 174-188.

Yates, J. F., Lee, J.-W., Shinotsuka, H., Patalano, A. L., & Sieck, W. R. (1998). Cross-cultural variations in probability judgment accuracy: Beyond general knowledge overconfidence? *Organizational Behavior and Human Decision Processes, 74*, 89-117.

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.