國立臺灣師範大學教育心理與輔導學系
教育心理學報，民91，34卷，1期，123—138頁

Sex Differences in Statistical Reasoning

HUI-JU C. LIU
Department of English Language
Da-Yeh University

JOAN B. GARFIELD
Department of Educational Psychology
University of Minnesota

Despite considerable research having been done in the area of sex differences in mathematical ability, statistical ability has rarely been the subject of a major research effort. This study focuses on the question of whether there are sex differences in statistical reasoning for college students. Participants included 245 college students in Taiwan and 267 American college students. The Statistical Reasoning Assessment (SRA) was used in this cross-cultural study to assess students' statistical reasoning ability. While the original version of the test was administered to students in the United States, a Chinese version of the instrument was administered to participants in Taiwan. Statistical methods were used to ascertain whether there were mean differences between males and females and whether there was equality between the correlation matrices for males and females. All the analyses are based on both the correct reasoning scores and the misconception scores obtained from the SRA instrument. Results tend to support the general research findings that when sex differences appear, they are in the direction favoring males, particularly in higher cognitive tasks such as mathematical reasoning. Analysis of the correlation matrices suggest that there are no differences in statistical reasoning between males and females for both countries. However, it should be noted that the results may be due to low item intercorrelations.

KEY WORDS: sex differences, statistical reasoning, misconception, correlation matrix

There has been considerable research on people's reasoning about probability over the past few decades. A wide variety of research studies in this area has focused on the errors made in probabilistic reasoning by children of all ages, college students and adults (Lecoutre, 1992; Kahneman & Tversky, 1972, 1973; Konold, 1989; Konold, Pollatsek, Well, Lohmeier, & Lipson, 1993; Tversky & Kahneman, 1971, 1974). Both mathematics educators and psychologists have contributed to the research in statistical reasoning. Psychologists are primarily concerned with the difficulties people have with reasoning about probability and statistics or even about everyday life problems under the situation of uncertainty. Mathematics and statistics educators are interested in the effect of instruction on helping students confront and correct their conceptual misunderstandings.

Despite considerable research on reasoning about probability and statistics, sex differences in

statistical reasoning have rarely been reported as major research findings. In the area of mathematical learning, however, sex differences have been the subject of research over many decades. Approximately 10% of the articles published in one major mathematics education journal, the Journal for Research in Mathematics Education, during the twelve-year period 1978-1990, had a gender theme (Leder, 1992). The present study intends to ascertain whether there are differences in reasoning about probability and statistics between males and females in two countries, Taiwan and the United States.

The research on statistical reasoning has been greatly influenced by the work of Daniel Kahneman and Amos Tversky since the early 1970s. Research studies by Kahneman and Tversky via the "heuristics and biases" approach have shown that people employ a limited number of heuristics when making predictions and judgments about probability under uncertainty (Kahneman & Tversky, 1973; Tversky & Kahneman, 1974). These heuristic principles help people reduce the complexity of tasks involving assessing probabilities and often result in quick and reasonable judgments, but sometimes they may lead to "severe and systematic errors" that are at odds with probability theory. Tversky and Kahneman (1971) suggested that even people with statistical training are prone to the same heuristics and biases as naive subjects. They believe that these heuristics are also prevalent in our real life situations when making numerous decisions based on the likelihood of uncertain events (Kahneman & Tversky, 1972).

Researchers such as Nisbett, Krantz, Jepson, and Kunda (1983), who have extensively explored the effects of statistical training on subjects, found that people apply statistical heuristics in reasoning about everyday life problems. They suggested that the important determinants for people to apply statistical heuristics are (1) clarity of the sampling process and the sample space, (2) presence of the role of chance factors in producing events, and (3) cultural prescriptions to reason statistically. Konold (1989) argued that some people did not reason probabilistically through judgment heuristics or via formal probability theory, but according to an outcome approach. He suggested that outcome-oriented people perceive each trial in an experiment as an individual event and their task is to successfully predict the outcome of the next single trial rather than to estimate the likelihood of its occurrence over the course of many trials.

Nisbett (1983) and his colleagues (Fong, Krantz, & Nisbett, 1986; Jepson, Krantz, & Nisbett, 1983) maintained that by offering formal training on statistical rules, people's ability to reason statistically can be improved. The same notion about the effectiveness of instruction is also supported by the work of Fischbein (1975) and his colleagues (Fischbein & Gazit, 1984; Fischbein, Pampu, & Manzat, 1970a, 1970b) and Kosonen and Winne (1995). However, other researchers including Garfield and Ahlgren (1988), Mevarech (1983), Shaughnessy (1977), and Well, Pollatsek, and Boyce (1990) suggested that misconceptions are difficult to overcome and change. They believe that certain misconceptions are deeply rooted in students' thinking and may not be overcome by mere exposure to statistics courses. Konold et al. (1993) claimed that it is important for educators to become familiar with the variety of students' misconceptions before they can try to change them. Garfield (1994) suggested that students' informal understanding and knowledge of a particular statistics concept needs to be challenged before formal instruction can become effective.

Although there is considerable research on statistical reasoning, sex differences in statistical reasoning have rarely been analyzed in previous research. The literature revealed considerable research on mathematical ability, typically showing sex differences favoring males (Benbow & Stanley, 1980, 1983; Dye & Very, 1968; Hilton & Berglund, 1974; Maccoby & Jacklin, 1974; Very, 1967; Very & Iacono, 1970). However, it is important to note that sex differences may depend on the portion of the distribution that is being studied, including age of the subjects and the complexity of the tasks tested. Most researchers suggest that there are no significant sex differences in mathematical learning before the seventh or eighth grades (Maccoby & Jacklin, 1974; Fennema, 1977). Differences start to appear after the elementary school years, but they do not always appear. A general consensus is that if differences appear, females tend to score higher than males on tasks that involve less cognitive complexity, such as rote memory, arithmetic computation, and verbal tasks. Males, on the other hand, tend to outperform females in tasks involving higher cognitive complexity, such as mathematical reasoning, problem solving, and tasks that require visual-spatial ability. Researchers such as Benbow and Stanley (1980, 1983) and Stevenson, Hale, Klein, and Miller (1968) found that sex differences also depend on the ability level of the subjects. In Benbow and Stanley's large-scale study of intellectually talented youth, sex differences in mathematical reasoning are found particularly noticeable for students who are at the higher end of the ability distribution. Other factors that are associated with sex differences include cognitive (i.e. verbal ability, spatial ability), affective (i.e. attitudes toward math, stereotyping math as a male domain, achievement motivation in math) and educational variables (i.e. course taking, teachers, school organizations). The question of sex differences is of multi-faced nature, broad and complex.

Traditional assessment of statistical knowledge rarely provides information about how students apply their probability and statistics knowledge to reason and solve problems (Garfield, 1998). The Statistical Reasoning Assessment (SRA) used in the present study is the first instrument developed to measure students' reasoning skills and misconceptions of statistics and probability. Items from this instrument were either adapted from or inspired by previous research studies (Liu, 1998). The main research question in the study was whether there are sex differences in reasoning about probability and statistics for college students in the two countries, Taiwan and the United States. Besides merely examining mean differences between the sexes, the equality of correlation matrices for males and females is also investigated by country.

Method

Subjects

Samples of this study included 94 males (38.4%) and 151 females (61.6%) from two universities in Taiwan, and 152 males (56.9%) and 115 females (43.1%) from one large midwestern university in the United States. The 245 students in Taiwan were majoring either in

Information Management or International Trade at the time of the test. They were mostly sophomores. The majority of the 267 students in the United States were either freshmen or sophomores majoring in business. All the subjects were at the end of an introductory business statistics course when the test was administered in the 1995-1996 academic year.

The Instrument

The instrument used in the study is the Statistical Reasoning Assessment (SRA; Liu, 1998), a 20 multiple-choice test. It was the first paper-and-pencil instrument developed to assess students' statistical reasoning, which students do not usually learn from the traditional curriculum. A test-retest reliability of .70 for correct reasoning items and .75 for misconception items of the instrument was obtained. The only evidence in support of validity of this instrument is content validity, which is mainly based on subjective judgments from experts. Factor analysis of the item intercorrelations failed to provide evidence of content validity for this instrument due to small inter-relationships of the items. There is also a lack of criterion-related validity for the instrument since no other measure of the same construct has been found to correlate with the test scores.

Items from the instrument were designed to measure students' correct reasoning skills and misconceptions respectively in eight different areas. For each item, there may be one single correct response or multiple correct responses. An item may measure one or more than one correct reasoning skill or misconception. For instance, there are three alternatives for item 12. Alternative A measures misconception involving law of small numbers. Alternative B measures whether examinees correctly understand the importance of large samples. Alternative C measures outcome orientation misconception. Thus, three item scores, one for the correct conception (12b) and two for different misconceptions (12a and 12c), are derived from item 12. It should be noted that not all the items measures both correct reasoning skill and misconception. Item 7 only measures misconception when alternative B or C is selected. No alternative was designed to measure correct reasoning skill for item 7.

Rather than obtaining 20 individual item scores and a total composite score depending on whether an examinee responds correctly, each examinee therefore has 19 item scores, eight subscale scores and one total composite score for his correct reasoning, and 21 item scores, eight subscale scores and one total composite score for his misconceptions. Table 1 shows the eight correct reasoning scales and eight misconception scales and the corresponding items and alternatives designed to measure each conception and misconception.

The study called for an equivalent Chinese version of this instrument. Unless a valid translated version of the instrument is used, the validity of any interpretation of the test results can be questioned. The instrument was translated into Chinese 1995 to be given to college students in Taiwan. Considerable efforts were made in the translation process to identify problems associated with translation. A procedure called "back translation" was used as an initial check of translation equivalence (Brislin, 1970). The translated Chinese version was later revised many times after being reviewed by people from both cultures, with or without statistics background, to ensure that the translation was native-sounding the original meaning was understood clearly and correctly.

Table 1 Correct Reasoning Skills and Misconceptions Measured by the SRA and the Corresponding Items and Alternatives for Measuring Each Conception and Misconception

Correct Reasoning Skills	*Corresponding Items and Alternatives*
1. Correctly interprets probabilities	2d, 3d
2. Understands how to select an appropriate average	1d, 4ab, 17c
3. Correctly computes probability	8c, 13a, 18b, 19a, 20b
4. Understands independence	9e, 10df, 11e
5. Understands sampling variability	14b, 15d
6. Distinguishes between correlation and causation	16c
7. Correctly interprets two-way tables	51d*
8. Understands importance of large samples	6b, 12b
Misconceptions	
1. Misconceptions involving averages	1a, 17e, 1c, 15bf, 17a
2. Outcome orientation misconception	2e, 3ab, 11abd, 12c, 13b
3. Good samples have to represent a high percentage of the population	7bc, 16ad
4. Law of small numbers	12a, 14c
5. Representativeness misconception	9abd, 10e, 11c
6. Correlation implies causation	16be
7. Equiprobability bias	13c, 18a, 19d, 20d
8. Groups can only be compared if they are the same size	6a

*Note: For item 5, subjects have to choose from two options before they can make further selection from four alternatives under each option. Alternative D under option 1 for item 5 measures whether examinees correctly interpret two-way tables.

Data Analysis

To ascertain whether there were mean differences in statistical reasoning between the sexes for college students, both the differences in the total correct reasoning scores and the total misconception scores for males versus females were tested using two-way analysis of variance (ANOVA). A two-way ANOVA of the total scores by sex and country was used to study the effects of sex, country, and the interaction between sex and country.

Equality between the correlation matrices for males and females were also analyzed. Pearson product-moment correlation matrices based on the correct reasoning item scores and misconception item scores are respectively obtained for each sex and split-sex group. The purpose of splitting each sex group in half is to examine discrepancies within a group. These within group discrepancies give some idea of the amount of expected sampling error and serve as a basis for interpreting differences between the sexes.

To ascertain the equality between the correlation matrices for males and females, two descriptive indices were obtained. One was the root-mean-square error term (RMSE) for the difference in two correlation matrices (Rock, Werts, & Flaugher, 1978). The RMSE value is

defined as the square root of the mean of the squared discrepancies of the corresponding elements of the two correlation matrices. If the two correlation matrices being compared are similar, the RMSE values will be small. The other was the percentage of discrepancies that falls below certain levels (i.e. .05, .10, .15, etc.) for the absolute difference of each corresponding value in the two correlation matrices (Lei & Skinner, 1982). A higher proportion of the discrepancies will be small if the two correlation matrices being compared are similar.

Results

Analysis of Mean Differences

The results for the two-way analysis of variance on the total correct reasoning scores by country and sex are presented in Table 3. Country effect appears to be highly significant (p<.01). Students in Taiwan have higher correct reasoning scores than their counterparts in the United States. Both the sex effect and the interaction effect between country and sex are not significant. However, both the effects are on the margin of being significant at the .05 level.

Table 2 Cell Means of Total Correct Reasoning Scores for Males and Females in Each Country

	Taiwan	United States	Total
Male	22.90 (4.83)	20.55 (4.59)	21.45
Female	21.38 (4.80)	20.57 (4.58)	21.03
Total	21.97	20.56	21.23

Note: Numbers in ( ) are standard deviations.

Table 3 Analysis of Variance Result for the Total Correct Reasoning Scores by Country and Sex

Source of Variation	SS	df	MS	*F*	F-prob
Country	307.27	1	307.27	13.95	<.001 **
Sex	69.05	1	69.05	3.13	077
Country x Sex	73.03	1	73.03	3.31	.069
Error	11193.60	508	22.04		

** $p < .01$

The ANOVA result for the total misconception scores by country and sex is presented in Table 5. Both country and sex effects appear significant ($p < .01$). Students in Taiwan have significantly lower misconception scores than students in the United States. Also, males show significantly lower misconception scores than their female counterparts.

Table 4 Cell Means of Total Misconception Scores for Males and Females in Each Country

	Taiwan	United States	Total
Mal	11.28 (4.42)	12.87 (4.05)	12.22
Female	12.81 (3.53)	13.39 (4.11)	13.09
Total	12.26	13.06	12.68

Note: Numbers in ( ) are standard deviations.

Table 5 Analysis of Variance Result for the Misconception Total Scores by Country and Sex

Source of Variation	SS	df	MS	*F*	F-prob
Country	145.42	1	145.42	9.13	.003 **
Sex	129.70	1	129.70	8.14	.005 **
Country x Sex	31.26	1	31.26	1.96	.162
Error	8095.00	508	15.94		

** $p < .01$

Although it is also the interest of the study to investigate sex differences within each culture, no further test was conducted in the first phase of the data analysis due to lack of significance for the interaction effect.

Analysis of the Equality of Correlation Matrices

Problems occur when creating correlation matrices that include items with extreme p-values since correlation depends on covariation. When there is no variability, there is no covariation and hence no correlation. Items with extremely low or high p-values were therefore removed from the analyses. Items deleted include five correct reasoning items (1d, 2d, 8c, 9e, and 12b) and seven misconception items (1ac, 2e, 9abd, 10e, 12a, 12c, 16ad).

The root-mean-square error term (RMSE) and the cumulative proportions of absolute discrepancies for the correlation matrices based on the correct reasoning scores are presented in Table 6. It is expected that the average intrasex RMSE value should be lower than the corresponding average intersex value for the split-sex groups, and there should be higher proportion of discrepancies for the within-sex groups than the between-sex groups below the same level since there should be more correspondence for the correlation matrices within sex than between the sexes. As shown in Table 6, the RMSE value for the Male-Female (0.143) is lower than the within split-sex values for the Male1-Male2 (0.195) and Female1-Female2 (0.161) of Taiwan samples. The increase in the RMSE values for these within-sex comparisons may result from the decrease in the sample size and the stability of the correlation matrices for the split-sex groups. The average RMSE value for the within-sex differences is 0.178 (the average of 0.195 for Male1-Male2 and 0.161 for Female1-Female2). The average between-sex RMSE value for the more comparable split-sex groups is 0.194 (the average of 0.209 for Male1-Female1, 0.188 for Male1-Female2, 0.195 for Male2-Female1, and 0.183 for Male2-Female2). Comparison of these

two average RMSE values (0.178 vs. 0.194) shows that there is a 9% increase in the between-sex RMSE value relative to the within-sex RMSE value. The comparison also shows more homogeneity within sex, as expected.

Table 6 Descriptive Indices for Discrepancy Between the Correlation Matrices for Males and Females in Taiwan and the United States Based on Correct Reasoning Scores

	RMSE	<.05	<.10	<.15	<.20	<.25	<.30	<.35	<.40	<.45	<.50
TAIWAN											
M-F	.143	.209	.440	.725	.824	.901	.956	.989	1.000	1.000	1.000
M1-M2	.195	.231	.374	.582	.703	.780	.923	.967	.978	.978	1.000
F1-F2	.161	.231	.484	.681	.802	.857	.934	.978	.989	.989	1.000
M1-F1	.209	.165	.451	.527	.681	.758	.835	.890	.956	.978	1.000
M1-F2	.188	.220	.407	.604	.758	.857	.901	.912	.956	.967	1.000
M2-F1	.195	.209	.385	.516	.637	.791	.868	.945	.978	.978	1.000
M2-F2	.183	.264	.352	.495	.758	.846	.890	.923	.989	.989	1.000
USA											
M-F	.136	.275	.538	.703	.846	.934	.967	1.000	1.000	1.000	1.000
M1-M2	.176	.253	.352	.615	.736	.912	.934	.945	.956	.967	1.000
F1-F2	.184	.209	.407	.593	.736	.824	.868	.934	.967	.967	1.000
M1-F1	.172	.264	.429	.604	.692	.879	.901	.967	.978	.978	1.000
M1-F2	.214	.209	.418	.604	.703	.791	.846	.890	.901	.923	1.000
M2-F1	.183	.286	.407	.604	.747	.813	.901	.934	.956	.978	1.000
M2-F2	.177	.220	.385	.560	.725	.791	.923	.956	.989	1.000	1.000

The cumulative proportions of absolute discrepancies less than 0.05 to 0.5 are given in steps of 0.05. M and F respectively represents the full sex samples for males and females. M1, M2, F1, and F2 represent the split-sex groups for males and females respectively.

The distribution of cumulative proportions of absolute discrepancies shows that an average of 87.9 percent of the discrepancies in the corresponding correlations for the between-sex groups (83.5% for Male1-Female1, 90.1% for Male1-Female2, 86.8% for Male2-Female1, and 89% for Male2-Female2) are less than 0.30, as compared with an average of 92.9 percent of the within-sex discrepancies (92.3% for Male1-Male2 and 93.4% for Female1-Female2). There is only a 5.5 percent decrease for the between-sex discrepancies relative to the corresponding value within sex. The discrepancies between the sexes are very small as contrasted with the intrasex differences.

Similarly, for students in the United States, there is only a 4% increase in the intersex RMSE value (0.187) as contrasted with the RMSE value within sex (0.180). Comparison of these split-sex groups show that within 0.30, there is only 0.8% decrease in the proportion of the between-sex discrepancies (89.3%) relative to the within-sex value (90.1%).

Table 7 Descriptive Indices for Discrepancy Between the Correlation Matrices for Males and Females in Taiwan and the United States Based on Misconception Scores

	RMSE	<.05	<.10	<.15	<.20	<.25	<.30	<.35	<.40	<.45	<.50
TAIWAN											
M-F	.142	.308	.571	.736	.835	.890	.956	.989	.989	.989	1.000
M1-M2	.217	.165	.429	.538	.670	.769	.824	.857	.912	.956	1.000
F1-F2	.148	.253	.473	.736	.835	.901	.945	.978	.978	1.000	1.000
M1-F1	.223	.143	.363	.527	.681	.758	.824	.879	.923	.934	1.000
M1-F2	.217	.198	.385	.560	.659	.758	.824	.868	.934	.956	1.000
M2-F1	.161	.308	.462	.670	.791	.879	.945	.967	.989	.989	1.000
M2-F2	.165	.242	.462	.670	.791	.879	.945	.967	.967	.978	1.000
USA											
M-F	.127	.286	.538	.769	.879	.956	.978	.989	1.000	1.000	1.000
M1-M2	.187	.176	.473	.648	.802	.846	.890	.923	.956	.967	1.000
F1-F2	.178	.198	.341	.571	.703	.835	.879	.978	1.000	1.000	1.000
M1-F1	.182	.198	.473	.582	.725	.868	.912	.934	.956	.978	1.000
M1-F2	.190	.253	.473	.604	.725	.813	.857	.879	.945	.989	1.000
M2-F1	.184	.165	.429	.582	.736	.824	.846	.945	.967	1.000	1.000
M2-F2	.164	.319	.484	.637	.725	.868	.923	.989	.989	.989	1.000

The cumulative proportions of absolute discrepancies less than 0.05 to 0.5 are given in steps of 0.05. M and F respectively represents the full sex samples for males and females. M1, M2, F1, and F2 represent the split-sex groups for males and females respectively.

Table 7 shows the RMSE and the cumulative proportions of absolute discrepancies in the corresponding correlations of the misconception scores. For the Taiwan samples, there is a 5% increase in the between-sex RMSE value (0.192) as compared with the within-sex value (0.183). For the samples in the United States, there is an unexpected 2% decrease in the between-sex RMSE value (0.180) in contrast to the corresponding within-sex RMSE value (0.183). Comparisons of these split-sex groups within 0.30 show that equal proportions of intersex discrepancies and intrasex discrepancies (88.5%) are obtained for samples in both countries.

If the two correlation matrices being compared are similar, there will be low RMSE values and higher proportion of small discrepancies of the corresponding values for pairs of matrices. The results show that the intersex RMSE values for the split-sex samples as contrasted with the corresponding intrasex RMSE values were either small or unexpectedly decreased. Also, decreases in the proportions of the intersex discrepancies for the split-sex samples are relatively small in contrast to the corresponding intrasex differences less than 0.30. There were even equal proportions of discrepancies between the sexes below 0.30 relative to the corresponding within split-sex values. These findings indicate that there are no differences in the correlation matrices between males and females for both countries.

Summary and Conclusion

The purpose of the present study is to ascertain whether there are sex differences in statistical reasoning. All the test results are based on examinees' correct reasoning scores and misconception scores obtained from the Statistical Reasoning Assessment (SRA), an instrument developed to assess students' understanding of probability and statistics concepts and reasoning skills.

The two-way analysis of variance test results based on both the correct reasoning scores and the misconception scores show that country effect is highly significant. Students in Taiwan have significantly higher correct reasoning scores and significantly lower misconception scores than students in the United States. Results also show that both sex effect and interaction effect between country and sex are nonsignificant when the total correct reasoning score is used as the dependent variable in the ANOVA test. However, both effects are on the margin of being significant at the .05 level. Plotting the cell means of the total correct reasoning scores for each sex by two countries shows that the lines for males and females are not parallel. The lines should be parallel if there is no interaction effect. The results suggest the strong possibility of interaction effect between country and sex. Males tend to have higher correct reasoning scores than females, while males and females in the United States have approximately equal performance on the same test items. When the total misconception score is used as the dependent variable, sex effect becomes highly significant. Males have lower total misconception scores than their female counterparts.

These results provide evidence in support of the general sex differences findings that when differences between the sexes appear, they tend to favor males, particularly on tasks involving higher level cognitive skills such as mathematical reasoning and problem solving. However, it is essential to understand that sex differences are a function of a combination of differential factors, including cognitive, affective and educational factors, rather than a function of simple factors. Sex differences cannot be considered as inferiority of either sex. Various socialization factors as well as biological factors may be involved in determining the differences between males and females. The stereotyping of mathematics as a male domain, the perceived attitudes of significant others, confidence in learning mathematics, and attributions of success or failure in mathematics would all become factors related to sex differences in this area. Research shows that females are less motivated and encouraged to pursue mathematics coursework and related occupations than males (Fennema & Sherman, 1977). This has also been a common situation for students in Taiwan. Male students have long been expected to have superior performance in science and mathematics than females in Taiwan. As above-mentioned, males are more likely to pursue mathematics or science related occupations than females. Therefore, males are more likely to receive more intense mathematics training than their female counterparts in Taiwan. Research findings suggest that this social-cultural factor may be one of the sources that leads to greater discrepancies for male and female students in Taiwan than for students in the United States. Future studies that assess a multitude of factors related to sex differences are needed to provide insightful investigations for the research questions of interest.

Although the results in the present study indicate that students in Taiwan exceed students in

the United States in their performance on the SRA test, not much confidence can be placed in the generalizability of the findings. It is recommended that the study be replicated by selecting more schools throughout the two countries to enhance generalizability of the results in the future. Different colleges in Taiwan may have different admission standards. Researchers suggest that sex differences are likely to exist depending on the portion of the population that is being studied (Benbow & Stanley, 1983). Thus, more schools should be selected to represent the variability of the whole population of males and females in both cultures. It will be interesting to see if replications of this study will yield similar results. It is also of interest to investigate whether sex differences are consistent across the countries in future studies.

Further, considerable efforts were made in the process of test translation. Items with problems associated with translation were identified either by using the back translation technique or by having different people in both the United States and Taiwan review the translated version of the instrument. However, researchers such as Hambleton and Bollwark (1991) suggested that, to test the equivalence of source and target versions, a combination of judgmental methods and empirical methods should be used. Errors missed by one method may be identified by another method. Statistical technique is also needed to verify translation quality. It is therefore recommended that in future studies, the schedule and budget of a cross-cultural study should provide for the time and money necessary to deal with the issue of scale comparability. Statistical analyses such as examination of the mean score differences, correlation coefficients, and equivalence of the factor structures of the two language forms should be used in combination with judgmental methods in the future to establish item equivalence.

Lastly, results based on the comparisons of the RMSE values and correlation discrepancies for the split-sex groups suggest that there are no sex differences in the correlation matrices for males and females in both countries. As foretold, if there is similarity between the correlation matrices for males and females, both the RMSE value and a higher proportion of the discrepancies will be small. Findings show that when there are increases in the between-sex RMSE values and decreases in the proportion of the intersex discrepancies as compared with the corresponding within-sex values for all the comparable split-sex samples, these increases and decreases are relatively small. However, it should be noted that most items from the instrument have very low item intercorrelations. Only three correct reasoning items (18b, 19a, and 20b), that measure combinatorial reasoning, and three misconception items (18a, 19d, and 20d), that measure equiprobability bias, have higher item intercorrelations. The SRA is designed to measure eight different correct reasoning skills and misconceptions. It is different from traditional assessment of probability and statistics knowledge, which relies heavily on numerical computation and has only one total score. Therefore, one should be very cautious when interpreting the results since small RMSE values and low discrepancies in the corresponding correlations may be due to low item intercorrelations.

References

Benbow, C. P., & Stanley, J. C. (1980). Sex differences in mathematical ability: Fact or artifact? *Science, 210*, 1262-1264.

Benbow, C. P., & Stanley, J. C. (1983). Sex differences in mathematical reasoning ability: more facts. *Science, 222*, 1029-1031.

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

Dye, N. W., & Very, P. S. (1968). Growth changes in factorial structure by age and sex. Genetic *Psychology Monographs, 78*, 55-88.

Fennema, E. (1977). Influences of selected cognitive, affective, and educational variables on sex-related differences in mathematics learning and studying. In L. H. Fox, E. Fennema, & J. Sherman (Eds.), *Women and mathematics: research perspectives for change* (pp.79-135). Washington, D. C.: National Institute of Education.

Fennema, E. H., & Sherman, J. A. (1977). Sex-related differences in mathematics achievement, spatial visualization and affective factors. *American Educational Research Journal, 14* (1), 51-71.

Fischbein, E. (1975). *The intuitive sources of probabilistic thinking in children.* Dordrecht, The Netherlands: Reidel.

Fischbein, E., & Gazit, A. (1984). Does the teaching of probability improve probabilistic intuitions? *Educational Studies in Mathematics, 15*, 1-24.

Fischbein, E., Pampu, I., & Manzat, I. (1970a). Comparison of ratios and the chance concept in children. *Child Development, 41*, 377-389.

Fischbein, E., Pampu, I., & Manzat, I. (1970b). Effects of age and instruction on combinatory ability in children. *The British Journal of Educational Psychology, 40*, 261-270.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistics training on thinking about everyday problems. *Cognitive Psychology, 18*, 253-292.

Garfield, J. B. (1994). Informal and formal conceptions of statistical power. Paper presented at the Fourth International Conference on Teaching Statistics, Morocco.

Garfield, J. B. (1998). Challenges in assessing statistical reasoning. Paper presented at the AERA 1998 annual meeting, San Diego.

Garfield, J. B., & Ahlgren, A. (1988). Difficulties in learning basic concepts in probability and statistics: implications for research. *Journal for Research in Mathematics Education, 19* (1), 44-63.

Hambleton, R. K., & Bollwark, J. (1991). Adapting tests for use in different cultures: technical issues and methods. *Bulletin of the International Test Commission, 18*, 3-32.

Hilton, T. L., & Berglund, G. W. (1974). Sex differences in mathematics achievement- a longitudinal study. *The Journal of Educational Research, 67* (5), 231-237.

Jepson, C., Krantz, D. H., & Nisbett, R. E. (1983). Inductive reasoning: competence or skill? *The Behavioral and Brain Sciences, 6*, 494-498.

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

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237-251.

Konold, C. (1989). Informal conceptions of probability. *Cognition and Instruction, 61* (1), 59-98.

Konold, C., Pollatsek, A., Well, A., Lohmeier, J., & Lipson, A. (1993). Inconsistencies in students' reasoning about probability. *Journal for Research in Mathematics Education, 24* (5), 392-414.

Kosonen, P., & Winne, P. H. (1995). Effects of teaching statistical laws on reasoning about everyday problems. *Journal of Educational Psychology, 87* (1), 33-46.

Leder, G. C. (1992). Mathematics and gender: changing perspectives. In D. Grouws (Ed.), *Handbook for research on mathematics teaching and learning* (pp.597-622). New York: Macmillan.

Lecoutre, M. P. (1992). Cognitive models and problem spaces in "purely random" intuitions. *Educational Studies in Mathematics, 23*, 557-568.

Lei, H., & Skinner, H. A. (1982). What differences does language make? Structural analyses of the personality research form. *Multivariate Behavior Research, 17*, 33-46.

Liu, H. J. (1998). *A cross-cultural study of sex differences in statistical reasoning for college students in Taiwan and the United States.* Doctoral dissertation, University of Minnesota, Minneapolis.

Maccoby, E. E., & Jacklin, C. N. (1974). *Psychology of sex differences.* Palo Alto, CA: Stanford University Press.

Mevarech, Z. R. (1983). A deep structure model of students' statistical misconceptions. *Educational Studies in Mathematics, 14*, 415-429.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday reasoning. *Psychological Review, 90*, 339-363.

Rock, D. A., Werts, C. E., & Flaugher, L. (1978). The use of analysis of covariance structures for comparing the psychometric properties of multiple variables across populations. *Multivariate Behavioral Research, 13*, 403-418.

Shaughnessy, J. M. (1977). Misconceptions of probability: an experiment with a small-group, activity-based, model building approach to introductory probability at the college level. *Educational Studies in Mathematics, 8*, 295-316.

Stevenson, H. W., Hale, G. A., Klein, R. E., & Miller, L. K. (1968). Interrelations and correlates in children learning and problem solving. *Monographs of the Society for Research in Child Development, 33* (7), 1-65.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*, 105-110.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science, 185*, 1124-1131.

Very, P. S., (1967). Differential factor structures in mathematical ability. *Genetic Psychology Monographs, 75*, 169-207.

Very, P. S. & Iacono, C. H. (1970). Differential factor structure of seventh grade students. *Journal of Genetic Psychology, 117*, 239-251.

Well, A. D., Pollatsek, A., & Boyce, S. J. (1990). Understanding the effects of sample size on the variability of the mean. *Organizational Behavior and Human Decision Processes, 47*, 289-312.

收 稿 日 期：2001年6月12日
接受刊登日期：2002年6月18日

Bulletin of Educational Psychology, 2002, 34(1), 123-138
National Taiwan Normal University, Taipei, Taiwan, R.O.C.

統計推理的性別差異

劉 慧 如
大葉大學
英美語文學系

JOAN B. GARFIELD
明尼蘇達大學
教育心理學系

摘 要

過去，在數學能力方面的性別差異（sex differences in mathematical ability）或在統計推理（statistical reasoning）兩個重要領域，都分別有過許多不同的研究，但是，卻不曾有人以統計推理方面的性別差異（sex differences in statistical reasoning）做爲研究主題。本研究主要在探討大學男女生在統計推理方面是否有性別差異。研究樣本包括245位台灣的大學生，以及267位美國的大學生。研究使用的問卷是統計推理測量（the Statistical Reasoning Assessment，SRA），在美國施測所使用的是英文原版問卷，在台灣施測時則使用翻譯成中文的版本。

本跨國研究使用不同的統計方法研討下列兩個主要問題：（1）男女生在統計推理方面是否有平均數的差異？（2）男女生之間的相關矩陣（correlation matrices）是否有所差異？所有的統計分析都是根據學生在回答問卷之後所得到的兩個分數：正確推理（correct reasoning）所得的分數及對於機率與統計不正確的觀念（misconception）所得的分數。針對第一個問題，研究結果偏向支持一般性別差異研究的發現：當男女生有性別差異時，尤其在較高層次認知的運用如數學推理方面，男生表現比女生優異。針對第二個研究問題所作的統計分析，結果顯示男生與女生的相關矩陣沒有差異。值得注意的是針對第二問題所得到的研究結果極可能導因於問卷（SRA）題目之間的相關係數（intercorrelations）都很小，因而導致使用相關矩陣來作資料分析的難題。

關鍵詞：性別差異、統計推理、迷思概念、相關矩陣