A Replication Study of Item Selection for the Bem Sex Role Inventory

Allen L. Edwards and Clark D. Ashworth
The University of Washington

An attempt was made to replicate the selection of items for the *Bem Sex Role Inventory (BSRI).* The 20 masculine, 20 feminine, and 20 neutral items of the *BSRI* were rated for social desirability "in an American male" and "in an American female" by male and female judges. The *BSRI* item-selection criterion—each item being rated by both male and female judges as significantly more desirable in a male than in a female (masculine items) or significantly more desirable in a female than in a male (feminine items)—was met by only two items: *masculine* and *feminine.* For a considerable number of other items, differences between mean desirability ratings for a male and for a female were in a direction opposite to that predicted. Correlations between the mean ratings of male and female judges when rating items for the same sex were quite high, consistent with previous research.

The *Bem Sex Role Inventory (BSRI),* developed by Bem (1974), consists of three scales of 20 items each. The *Masculinity* scale contains traits judged by both male and female judges to be significantly more desirable in an American male than in an American female, and the *Femininity* scale contains traits judged by both male and female judges to be significantly more desirable in an American female than in an American male. The items in the *Social Desirability* scale were not judged to be significantly more desirable in one sex than in the other by both male and female judges and are referred to by Bem as *neutral* items.

Bem obtained her ratings of the desirability of the items in the *BSRI* from undergraduate student judges in the winter of 1972 ($N = 40$) and the following summer ($N = 60$). Since students' conceptions of sex roles (and sex-role stereotypes) are undergoing rapid change on university campuses, items that were rated as significantly more desirable in one sex than in the other in 1972 might no longer be so rated. The present study was, therefore, undertaken to see whether Bem's item selection could be replicated.

Method

Two male and two female upper division undergraduate students doing supervised research for credit collected the ratings of social desirability for the 60 items in the *BSRI.*[1] Each experimenter was asked to obtain ratings of the desirability of each trait in an American male from 10 male and 10 female judges and ratings of the desirability of the traits in an American female from 10 male and 10 female judges. All of the ratings of social desirability were collected

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 1, No. 4 Fall 1977 pp. 501-507

[1]The four experimenters were Susan Wong, Patricia Pedigo, Douglas Edwards, and Mark Soelling.

on the same day in the fall of 1975. One of the male experimenters went to one campus library and one of the female experimenters went to another campus library to collect their ratings. The other male experimenter went to one student union and the other female experimenter went to a different union building to obtain their ratings.

Each experimenter had practice in approaching potential raters and asking if they would be willing to volunteer from five to seven minutes of their time to rate some personality traits for desirability. Each potential volunteer was told that he/she would be provided with a written explanation of the nature of the research on completing the ratings, and this was done. They were also assured that there was no need to identify their ratings with their names. When a rater volunteered, as did virtually all who were approached, the rater was provided with a description of a 9-point rating scale. The rating scale was the one used by Edwards (1970), with the instructions modified to be specifically for either an American male or an American female. For example, one set of instructions was entitled *Ratings of the Desirability of Traits in American Males* and the phrase "in an American male" was repeated three times in the detailed instructions. A similar set of instructions was prepared to obtain ratings of the desirability of the traits in an American female. Raters were also instructed that: "We are not interested in whether these personality traits do or do not describe you. We would simply like to know how desirable you judge them to be in an American male (female)."

Results

Overall Mean Ratings

Table 1 shows the completed returns obtained by each of the four experimenters. For example, Experimenter 1, a female, obtained 10 records of males rating a male *(MM)*, 10 records of males rating a female *(MF)*, 9 records of females rating a male *(FM)*, and 9 records of females rating a female *(FF)*. The overall mean ratings of social desirability obtained by the four experimenters for the 60 items in the *BSRI* were 6.17, 6.05, 6.08, and 6.00 respectively. These means do not differ significantly ($F = .11$; $df = 3, 236$).

Table 1

Number of raters obtained by each experimenter for each combination of sex of rater and sex rated and overall mean ratings of the social desirability of the 60 items in the BSRI

Experimenters	Sex of rater and sex rated					Mean Rating	s
	MM	MF	FM	FF	Total		
1. Female	10	10	9	9	38	6.17	1.71
2. Female	10	10	10	10	40	6.05	1.58
3. Male	10	10	10	10	40	6.08	1.70
4. Male	8	10	10	9	37	6.00	1.62
Total	38	40	39	38	155		
Mean rating	5.94	6.02	6.24	6.11		6.08	
s	1.53	1.76	1.76	1.74			

Masculine Items

According to Bem's results, masculine items should be rated significantly higher in desirability for a male than for a female by both male and female judges. For the male judges, 9 of the 20 masculine items had differences between the means in the predicted direction, but only two of the mean differences were significant with $\alpha = .05$ for a one-sided test.[2] The two significant items were *masculine* and *dominant*. The t ratios for the other 7 items were all less than 1.43. For each of the remaining 11 items, the direction of the difference between the means was reversed; that is, the items were rated as being more desirable in a female than in a male by the male judges, but not significantly so.

For the female judges, only 4 of the 20 masculine items had mean differences in the predicted direction, and only one was significant. That item was *masculine*. Of the other 16 items, one of the mean differences was significant.[3] Female judges rated *self-sufficient* as being significantly more desirable in a female than in a male.

Thus, of the 20 items in the *Masculine* scale only one item, *masculine*, survived Bem's criterion of being rated as significantly more desirable in a male than in a female by both male and female judges. To determine whether or not at least one of the four experimenters obtained ratings that were more in accord with Bem's results for the masculine items, the t tests described above were repeated for each experimenter separately. For each of the four experimenters, only *masculine* was rated as significantly more desirable in a male than in a female by both male and female judges.

[2]Bem used a two-sided test with $\alpha = .05$ in selecting the items for her scales from a larger pool of items with no prior knowledge of the direction of the mean differences. Given the expected direction of the mean difference, as reported by Bem, a one-sided test with $\alpha = .05$ was used here. The use of the more "liberal" one-sided test is "biased" towards confirming Bem's item selection.

[3]Two-sided tests with $\alpha = .05$ were used for these cases because the direction of the differences between the means was opposite of that predicted.

Feminine Items

The differences between the means of 15 of the 20 feminine items were in the predicted direction when rated by male judges. Of these 15 differences, 8 were significant: *cheerful, affectionate, feminine, compassionate, warm, does not use harsh language, loves children,* and *gentle.* But when these same 20 items were rated for desirability in a male and in a female by female judges, only two of the items had a difference between the means in the predicted direction, and only one of the differences was significant: *feminine.* For those 18 feminine items which the female judges rated as more desirable in a male than in a female, five of the differences between the means were significant: *yielding, cheerful, affectionate, tender,* and *gentle.*

Again, only one of the 20 items in the *Feminine* scale met Bem's criterion of being rated as significantly more desirable in females than in males by both male and female judges. The one significant item was *feminine.* To see if the ratings of the feminine items obtained by any one of the four experimenters would result in the selection of items other than *feminine,* the t tests were repeated for each experimenter separately. For each of the four experimenters only *feminine* was judged to be significantly more desirable in a female than in a male by both male and female judges.

Neutral Items

Of the 20 neutral items, the male judges rated *secretive* and *conceited* as being significantly less desirable in a female than in a male and *likable* and *friendly* as being significantly more desirable in a female than in a male. For the female judges, one item had a significant mean difference: *jealous* was rated as less desirable in a female than in a male.

Differences in Scale Means

Table 2 gives the mean desirability ratings for the 20 masculine, 20 feminine, and 20 neutral

items for male and female judges and corresponds to Bem's Table 2. Contrary to Bem's findings, the male judges in our sample did not rate masculine items as significantly more desirable for males than for females (t = .91), nor did the female judges (t = −.06).

For the male judges, the difference between the means for the feminine items was in the predicted direction. The mean desirability rating of the feminine items was higher when rated for females than for males and the difference was significant (t = −2.38). For the female judges, the difference between the mean ratings of the feminine items was not in the predicted direction. Female judges rated the feminine items as being slightly more desirable in a male than in a female (t = .92). Consistent with Bem's results, the

Table 2

Mean social desirability ratings of the masculine, feminine, and neutral items

Items	Male judges			Female judges		
	Masculine items	Feminine items	Neutral items	Masculine items	Feminine items	Neutral items
For a man	6.37	5.83	5.61	6.61	6.30	5.81
For a woman	6.17	6.33	5.56	6.62	6.08	5.64
Difference	.20	-.50	.05	-.01	.22	.17
s_D	.22	.21	.11	.18	.24	.09
t_D	.91	-2.38	.45	-.06	.92	1.89

Table 3

Mean social desirability ratings of the masculine and feminine items for one's own sex

Items	Male judges	Female judges
	for a man	for a woman
Masculine	6.37	6.62
Feminine	5.83	6.08
Difference	.54	.54
s_D	.41	.46
t_D	1.32	1.17

means for the neutral items did not differ significantly ($t = .45$ and $t = 1.89$).

Table 3 shows the mean desirability ratings of the masculine and feminine items for persons of the same sex as the judges, and it corresponds to Bem's Table 3. Male judges rated masculine items as more desirable in a male than the feminine items, but the difference was not significant ($t = 1.32$). For female judges, the difference was not in the direction found by Bem. Instead, female judges rated the masculine items as being more desirable in a female than the feminine items. Again the difference was not significant ($t = 1.17$).

Correlations Between Mean Ratings of Social Desirability

There is considerable evidence that mean ratings of social desirability of personality traits obtained from male and female judges with respect to generalized others (sex unspecified) are highly correlated and stable (Edwards, 1970). On the other hand, there appears to be no evidence regarding the correlation between mean ratings of social desirability obtained from male and female judges with respect to a specific sex, either male or female.

Table 4 shows the correlations between the mean ratings of female and male judges, when both groups of judges rated a female and when both groups rated a male. The correlations are given separately for the 20 feminine, 20 masculine, and 20 neutral items and also for the complete set of 60 items. For example, when females rated the desirability of the feminine items for a female *(FF)* and males rated the desirability of the same items for a female *(MF)*, the correlation between the average ratings was .97. When the feminine items were rated for desirability in a male by male *(MM)* and female *(FM)* judges, the correlation was also .97. Table 4 also gives the means of the ratings and the standard deviations for each set of items for each rating condition. For example, when females rated the feminine items for desirability in a female *(FF)*, the mean rating for the 20 items was 6.08 and the standard deviation was 1.65.

Table 4

Means, standard deviations, and correlation coefficients between mean ratings of social desirability when male and female judges rated the same sex. The correlation coefficients are given separately for each set of 20 items and for the combined set of 60 items.

	Feminine items					Masculine items			
	FF	MF	MM	FM		FF	MF	MM	FM
$\bar{X}$	6.08	6.33	5.83	6.30	$\bar{X}$	6.62	6.17	6.37	6.61
s	1.65	1.60	1.57	1.88	s	1.15	1.29	.88	.98
r	.97		.97		r	.92		.91	
	Neutral items					All items			
	FF	MF	MM	FM		FF	MF	MM	FM
$\bar{X}$	5.64	5.56	5.61	5.81	$\bar{X}$	6.11	6.02	5.94	6.24
s	2.21	2.25	1.93	2.17	s	1.74	1.76	1.53	1.76
r	.99		.99		r	.96		.97	

The lowest correlation shown in Table 4 is for the masculine items, when rated for desirability in a male by male *(MM)* and by female *(FM)* judges *(r = .91)*. All of the values shown in the table are consistent with previous findings regarding the correlation between social desirability scale values based on ratings of male and female judges with respect to generalized others.

A Second Attempt to Replicate Bem's Findings

Because the results of the above study were in considerable disagreement with the earlier findings of Bem, it was repeated in the fall of 1976 using two experimenters, one male and one female.[4] Each experimenter obtained ratings of the desirability of the Bem items from 20 males and 20 females so that, for the two experimenters combined, 20 males rated the desirability of the items in a male, 20 males rated the desirability of the items in a female, 20 females rated the desirability of the items in a male, and 20 females rated the desirability of the items in a female.

Again in this second attempt to replicate Bem's results, the only item rated by both males and females as significantly more desirable in a male than in a female was *masculine*, and the only item rated by both males and females as significantly more desirable in a female was *feminine*.

The mean desirability value of each of the 60 items was determined by pooling the ratings over sex of experimenter, sex of the raters, and sex rated. Each of these means was, therefore, based on the ratings of 80 judges. The mean of the means was 6.13, with a standard deviation of 1.66. The corresponding values obtained in the fall of 1975 were 6.08 and 1.64, respectively. The correlation between the two sets of ratings, obtained approximately one year apart, was .99.

[4]The two experimenters were Sandra Johnson and Norman Dorpat.

Discussion

How are we to account for the fact that the only item in two independent samples judged by both males and females to be significantly more desirable in a male than in a female was *masculine* and the only item judged by both groups to be more desirable in a female than in a male was *feminine*? A plausible explanation is simply that college students' conception of the feminine and masculine sex roles and sex-role stereotypes has changed since the time Bem collected her ratings of social desirability. Several other explanations are possible:

1. *Lack of power:* It is possible that some of the mean differences that were in the same direction as Bem's would have been significant with an increased number of judges. But, there were 38 to 40 judges in each of our groups, whereas Bem used only 25. In addition, a more lenient test of significance with $\alpha = .05$ for a one-sided test was used in the present study, while Bem used a two-sided test with $\alpha = .05$. Furthermore, a considerable number of the differences between the means obtained here were in a direction opposite to those obtained by Bem.
2. *Type I errors:* Bem selected the 20 items for the *Masculine* scale and the 20 items for the *Feminine* scale from a larger pool of approximately 200 items. It is possible that some of the items she selected represent Type I errors, but the possibility that a Type I error would occur for a given item for *both* male and female judges is very small and particularly so for as many as 40 items.
3. *Differences in rating scales:* In this study a 9-point rating scale ranging from 1 *(Extremely undesirable)* through 5 *(Neutral)* to 9 *(Extremely desirable)* was used to collect ratings. This scale permits a judge to rate the degree of undesirability of traits as well as the degree of desirability. Bem, on the other hand, used a 7-point scale ranging from 1 *(Not at all desirable)* to 7 *(Extremely*

desirable) in obtaining her ratings. Bem's scale appears to be ambiguous because it is not clear how judges would use it in rating traits they consider to be of average desirability or those they consider to be undesirable. It is as if one were asked to rate the "tallness" of males on a scale ranging from 1 *(Not at all tall)* to 7 *(Extremely tall)*. Is a male who is of average height "not at all tall" and, if so, then what is a male who is below average in height? The difference between Bem's scale and the one used here may account, in part, for the failure to replicate her findings.

4. *SD response bias:* A possibility exists that a generalized bias in making social desirability ratings could have affected the results of this study. For example, one group of raters may have rated most of the 60 items as much more desirable for one or more of the experimenters. However, the fact that the overall means of the ratings collected by each experimenter did not differ significantly does not support this hypothesis.
5. *Differences in methods of data collection:* In the present study both male and female upper division undergraduate students were used to collect data and the ratings were obtained individually rather than from a group. Bem did not specify how she obtained her ratings. If she was the only experimenter and if she obtained her ratings from classes in which she was the instructor, this may have had some influence on the ratings she obtained.
6. *Differential sampling:* Another possible explanation of the results obtained here is differential sampling. For example, Bem's findings might have been duplicated if students enrolled in introductory psychology courses had been used or if students had been sampled from a private university similar to Stanford rather than a state university. This seems unlikely, however, in view of the many and varied studies showing a high degree of agreement between diverse groups with respect to ratings of social desirability (Edwards, 1970).

References

Bem, S. L. The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 1974, *42*, 153–162.

Edwards, A. L. *The measurement of personality traits by scales and inventories*. New York: Holt, Rinehart and Winston, 1970.

Author's Address

Allen L. Edwards, Department of Psychology NI-25, The University of Washington, Seattle, WA, 98105.