Cross-Validation of Item Selection on the Bem Sex Role Inventory

Hugh Walkup and Robert D. Abbott
University of Washington

Prompted by Edwards and Ashworth (1977) Bem's (1974) BSRI item selection strategy was reexamined, using her anchored rating scale and instructions. The results replicate Bem's for 18 of the 20 masculine items and 19 of the 20 feminine items. These results suggest that the failure to replicate, reported by Edwards and Ashworth, resulted from differences in the anchored rating scale and instructions. One masculine and three feminine BSRI items were found to violate Bem's assumption of relative desirability. Only half the neutral items were rated neutral with regard to sex by the judges.

The Bem Sex Role Inventory (BSRI; Bem, 1974, 1976, in press), has stimulated much research and discussion which has generally supported and refined the psychometric properties and utility of the BSRI in confirming Bem's hypotheses (see Bem, 1975, Bem & Lenny, 1976; Bem, Martyna, & Watson, 1976; Kelley & Worell, 1976; Ruble, Croke, Frieze, & Parsons, 1975; Spence, Helmreich, & Stapp, 1975; Strahan, 1975; Wakefield, Sasek, Friedman, & Bowden, 1976). However, a serious challenge to the BSRI has been recently reported.

In that study (Edwards & Ashworth, 1977) only two items met the BSRI item selection criteria: "masculine" and "feminine." Several Masculine (*M*) and Feminine (*F*) items were rated in the direction opposite that found by Bem, i.e., several *M* items were rated as more desirable in a female than in a male, and vice versa, by either male or female judges. In Edwards and Ashworth (1977), the overall ratings of the items on the *M* and *F* scales were not significantly different for either male or female raters under either set of instructions.

The instructions to judges and rating scale used in the Edwards and Ashworth study differed considerably from those used by Bem when she selected the BSRI items. Bem's instructions were,

> We are interested in American stereotypes of masculinity and femininity. On the following page you will be shown a large number of personality characteristics. We would like you to indicate how desirable it is in American society for a man to possess each of these characteristics. Note: We are not interested in your personal opinion of how desirable each of these characteristics is. Rather, we want your judgment of how our society evaluates each of these characteristics in a man.
>
> Example: healthy
> Mark a 7 if it is considered EXTREMELY DESIRABLE in America for a man to be healthy.
> Mark a 6 if it is considered VERY DE-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 2, No. 1 Winter 1978 pp. 63–71

SIRABLE in America for a man to be healthy.

Mark a 5 if it is considered QUITE DESIRABLE in America for a man to be healthy.

Mark a 4 if it is considered MODERATELY DESIRABLE in America for a man to be healthy.

Mark a 3 if it is considered SOMEWHAT DESIRABLE in America for a man to be healthy.

Mark a 2 if it is considered SLIGHTLY DESIRABLE in America for a man to be healthy.

Mark a 1 if it is considered NOT AT ALL DESIRABLE in America for a man to be healthy.

Edwards and Ashworth's instructions were to rate "how desirable or undesirable you judge [these personality traits] to be in an American male/female." Edwards and Ashworth's nine-point rating scale was anchored by (1) "extremely undesirable," (2) "strongly undesirable," (3) "moderately undesirable," (4) "mildly undesirable," (5) "neutral," (6) "mildly desirable," (7) "moderately desirable," (8) "strongly desirable," (9) "extremely desirable."

Thus, the two sets of instructions differed in their emphasis on society versus individual rating and in reference to sex (male/female) or gender (man/woman) of the person being rated. In addition, the response scales differed in length and content. Since Bem was selecting from a generated set of items thought to be positive in value, she did not use the "undesirable" markers which are used in Edwards and Ashworth's Social Desirability rating scales.

Edwards and Ashworth (1977) identified several possible explanations for their failure to replicate Bem's findings: (1) changes in college students' conception of sex roles and sex role stereotypes; (2) differences in Stanford and Washington students; (3) differences in the length of the scale; (4) differences in the data collection procedures; and (5) differences in the Power and Type I error rates of the two studies.

Method

In order to test these possible explanations of the discrepancy between Edwards and Ashworth's (1977) and Bem's (1974) findings and to examine the appropriateness of assumptions about the positive value of BSRI items, the BSRI item selection procedures were cross-validated, using Bem's original instructions and rating scales with University of Washington students.

Students in four Educational Psychology classes (no student was in more than one class) at the University of Washington were adminis-

Table 1
Means and Standard Deviations (SD) of the Ratings of BSRI Masculine (M), Feminine (F), and Neutral (N) Items by Male and Female Judges "For a man" and "For a woman"

	Male Judges			Female Judges		
	M	F	N	M	F	N
Instructions	Items	Items	Items	Items	Items	Items
"For a man"						
Mean	5.26	3.62	3.80	5.78	3.69	3.89
SD	.75	.83	.48	.53	.63	.45
"For a woman"						
Mean	3.53	5.20	4.14	3.25	5.34	4.20
SD	.88	.55	.36	.88	.65	.51

tered Bem's instructions. Eighteen males and 30 females rated "In American society how desirable is it for a *man* to be . . ." and 23 males and 58 females rated the BSRI items following Bem's instructions "In American society, how desirable is it for a *woman* to be . . .". All subjects rated the BSRI items on Bem's anchored rating scale. A male experimenter was employed in all rating groups.

Results

Table 1 shows the overall means and standard deviations of ratings of the BSRI items for male and female judges in the two instruction groups. Tables 2, 3, and 4 show the means and standard deviations of the Masculine, Feminine, and Neutral BSRI items, respectively, as rated "for a man" and "for a woman" by male and female judges. The mean ratings shown in Tables 2, 3, and 4 will also be useful for other investigators using the BSRI or subsets of its items. Unless otherwise specified, the term "significantly different" has been used to refer to outcomes which would be expected to occur by chance less than 5 times in 100 using two-tailed tests.

"For a Man" Instructions

Masculine items were judged overall as significantly more desirable than Feminine items by both male [$t(17)=5.59$, $p<.001$] and female [$t(29)=12.31$, $p<.001$] judges. Female judges rated the Masculine items overall as more desirable "for a man" [$t(46)=2.81$, $p<.01$] than male judges. There was no such difference for Feminine [$t(46)=-.31$] or Neutral [$t(46)=-.64$] items.

Female judges rated "defends own beliefs," "independent," "athletic," "assertive," "has leadership abilities," "self-sufficient," "tender," and "individualistic" as significantly more desirable "for a man" than did male judges. No characteristics were rated as significantly more desirable "for a man" by male judges than by female judges.

"For a Woman" Instructions

Feminine items were rated overall as significantly more desirable "for a woman" than Masculine items by both male [$t(22)=7.50$, $p<.001$] and female [$t(57)=12.25$, $p<.001$] judges. There were no significant differences between the male and female judges in their overall ratings of Masculine and Neutral items.

Male judges rated "competitive" as more desirable "for a woman" than female judges. There were no significant differences between male and female judges on any other item.

Comparison of "For a Man" and "For a Woman" Instructions

Female judges rated Masculine items overall as significantly more desirable "for a man" than "for a woman" [$t(86)=14.53$, $p<.001$], Feminine items overall as significantly more desirable "for a woman" than "for a man" [$t(86)=11.43$, $p<.001$], and Neutral items as overall more desirable "for a woman" than "for a man" [$t(86)=2.87$, $p<.01$].

Male judges rated Masculine items overall as significantly more desirable "for a man" than "for a woman" [$t(39)=6.69$, $p<.001$], Feminine items overall as more desirable "for a woman" than "for a man" [$t(39)=7.35$, $p<.001$], and Neutral items overall as more desirable "for a woman" than for a man [$t(39)=2.64$].

Female judges rated *each* Feminine item as significantly more desirable "for a woman" than "for a man," and *each* Masculine item as significantly more desirable "for a man" than "for a woman."

Male judges rated *each* Feminine item as significantly more desirable "for a woman" than "for a man" except "loyal" [$t(39)=1.42$]. They rated *each* Masculine item as significantly more desirable "for a man" than "for a woman" except "willing to take risks" [$t(39)=.84$] and "individualistic" [$t(39)=.62$]. The exceptions were in the predicted direction, although they did not reach significance.

Table 2

Means and Standard Deviations (SD) of Ratings for Each Masculine BSRI Item

	"For a man"				"For a woman"			
	Male Judges		Female Judges		Male Judges		Female Judges	
Item	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Self-reliant[1]	6.17	1.04	6.50	.68	4.09	1.38	4.03	1.27
Defends own beliefs	5.33	1.19	6.07	.94	4.39	1.37	3.97	1.31
Independent	5.61	1.33	6.40	1.10	4.22	1.24	3.76	1.44
Athletic[1]	4.78	1.51	5.57	.86	3.65	1.37	3.66	1.40
Assertive[1]	5.00	1.14	6.00	.83	3.39	1.31	3.28	1.40
Strong personality	5.44	1.10	5.80	.76	4.22	1.45	3.57	1.54
Forceful	4.39	1.34	5.07	1.48	2.57	1.16	2.24	1.07
Analytical	4.72	1.18	5.10	1.27	3.61	1.44	3.09	1.34
Has leadership abilities[1]	5.61	1.33	6.30	.75	3.65	1.34	3.53	1.33
Willing to take risks	4.17	1.47	4.77	1.50	3.78	1.45	3.59	1.31
Makes decisions easily	5.61	.92	5.37	1.30	3.96	1.30	4.00	1.31
Self-sufficient[1]	5.56	1.72	6.47	.97	3.87	1.18	3.83	1.17
Dominant	4.72	1.57	4.97	1.61	1.91	1.04	1.67	1.00
Masculine	6.28	.83	6.47	.82	1.43	.79	1.34	.66
Willing to take a stand	5.72	1.23	6.07	.74	3.91	1.13	3.71	1.31
Aggressive	4.78	1.26	5.07	1.53	2.39	1.16	2.33	1.29
Acts as a leader	5.28	1.41	5.77	1.04	3.39	1.31	3.12	1.22
Individualistic[1]	4.61	1.58	5.80	.76	4.35	1.15	3.84	1.42
Competitive	5.39	1.04	5.93	1.11	3.52	1.24	2.88	1.27
Ambitious	6.11	.90	6.20	1.06	4.17	1.34	3.53	1.47

[1]The variance of ratings for males and females rating "for a man" differ dignificantly. Separate variance estimates were used in computing *t* values.

Table 3

Means and Standard Deviations (SD) of Ratings for Each Feminine BSRI Item

	"For a man"				"For a woman"			
	Male Judges		Female Judges		Male Judges		Female Judges	
Item	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Yielding	3.39	1.50	2.80	1.45	4.39	1.03	4.57	1.39
Cheerful	5.17	1.25	4.87	.90	5.87	.92	6.14	.91
Shy	1.89	.83	1.90	1.06	3.35	1.07	3.28	1.44
Affectionate	4.39	1.61	4.50	1.16	6.17	.78	6.24	.78
Flatterable	2.89	1.60	2.93	1.17	4.52	1.41	4.52	1.51
Loyal[1]	5.89	1.32	5.37	1.43	6.39	.94	6.19	1.02
Feminine	1.22	.94	1.40	1.30	6.26	.74	6.16	1.14
Sympathetic	3.94	1.43	4.23	1.07	5.74	1.01	6.10	.87
Sensitive to needs of others	4.28	1.41	5.00	1.39	5.91	1.08	6.10	.87
Understanding	4.94	1.51	5.03	1.10	6.04	.93	6.17	.75
Compassionate	4.17	1.51	4.70	.99	5.87	.97	6.07	.92
Eager to soothe hurt feelings	3.83	1.47	3.90	1.00	5.74	.96	5.78	.96
Soft spoken	3.22	1.52	3.00	1.20	4.13	1.10	4.47	1.45
Warm	4.28	1.32	4.60	1.10	5.87	.92	6.05	.98
Tender	3.28	1.53	4.27	1.26	5.87	.69	5.95	.91
Gullible	1.67	.91	1.23	.63	2.43	1.27	2.91	1.56
Childlike	1.39	.98	1.63	.85	2.39	1.50	2.95	1.55
Does not use harsh language	3.72	1.41	3.30	1.15	5.04	1.67	5.09	1.65
Loves children	4.72	1.27	4.80	1.13	6.04	.88	6.00	1.23
Gentle	4.17	1.34	4.30	1.09	6.00	.67	5.98	.96

[1]The variance of ratings for males and females rating "for a woman" differ significantly. Separate variance estimates were used in computing *t* values.

Table 4

Means and Standard Deviations (SD) of Ratings for Each Neutral BSRI Item

	"For a man"				"For a woman"			
	Male Judges		Female Judges		Male Judges		Female Judges	
Item	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Helpful	5.06	1.35	4.77	1.40	6.00	.80	5.93	.84
Moody	1.50	.86	1.63	.93	2.00	.91	1.91	1.23
Conscientious	5.50	1.34	5.43	1.38	5.26	1.05	5.50	1.20
Theatrical	2.33	1.19	2.37	1.00	3.35	1.30	3.03	1.53
Happy	5.22	1.26	5.20	1.19	5.91	.73	6.10	.87
Unpredictable	1.94	1.06	2.20	1.06	2.61	1.44	2.95	1.47
Reliable	6.00	1.09	6.10	.92	5.65	.89	5.78	.94
Jealous	2.94	1.83	2.97	1.77	3.00	1.32	2.72	1.73
Truthful	5.11	1.23	5.43	1.25	5.87	1.36	5.53	1.11
Secretive	2.39	1.15	2.47	1.25	2.52	1.16	2.81	1.36
Sincere	4.89	1.45	5.37	1.13	5.87	1.06	5.91	.90
Conceited	1.88	1.30	2.20	1.50	1.70	1.02	1.86	1.26
Likeable	5.44	.92	5.60	1.16	5.78	.90	6.09	.84
Solemn	3.00	1.46	3.43	1.31	2.91	1.62	2.57	1.24
Friendly	5.00	1.14	5.47	.90	5.61	.72	5.93	.90
Inefficient	1.28	.58	1.17	.53	1.57	.90	1.91	1.43
Adaptable	5.17	1.30	5.23	1.01	5.43	.99	5.33	.94
Unsystematic[1]	2.28	1.45	1.83	.79	2.13	.76	2.66	1.31
Tactful	4.83	1.20	4.90	1.09	5.00	.95	5.31	1.11
Conventional[1,2]	4.22	1.83	3.93	1.02	4.65	1.19	4.24	1.62

[1]The variance of ratings for males and females rating "for a man" differ significantly. Separate variance estimates were used in computing *t* values.

[2]The variance of ratings for males and females rating "for a woman" differ significantly. Separate variance estimates were used in computing *t* values.

Male and female judges rated the following Neutral items as significantly more desirable "for a woman" than "for a man": "helpful," "theatrical," "happy," "sincere," and "friendly." Female judges, only, also rated the Neutral items "unpredictable," "likeable," "inefficient," and "unsystematic" as significantly more desirable "for a woman" than "for a man" and the Neutral item "solemn" as significantly more desirable "for a man" than "for a woman."

Discussion

Using Bem's instructions and anchored rating scale, University of Washington judges closely approximated the mean desirability ratings reported by Bem (personal communication). Correlations between Bem's and the University of Washington judges' mean ratings of masculine, feminine, and neutral items ranged from .93 to .98 for paired rating groups. The judges' mean ratings of 19 of the 20 Feminine items and 18 of the 20 Masculine items met the BSRI item selection criteria. The three items which did not meet the criteria, "loyal," "willing to take risks," and "individualistic" were rated significantly different "for a man" than "for a woman" in the appropriate direction by female judges, but not by male judges.

It may be, as several feminist authors have suggested, that males are less sensitive to sex role stereotypes. Six of the Masculine items, as well as the overall Masculine scale, received ratings significantly higher from female judges than from male judges under "for a man" instructions. In light of male and female judges' agreement under "for a woman" instructions, this suggests an alternative hypothesis that disagreement among men and women remains on the relative desirability of sex role characteristics in men.

Bem's rating scale, then, appears most appropriate for detecting differences in desirability ratings of BSRI items. Are those detected differences meaningful? In an absolute sense, no. They do not tell us that an item is desirable for a woman and undesirable for a man, or vice versa; but that is not what BSRI set out to accomplish. Rather, the relative degree of desirability for women and men was the issue. While "acts as a leader" or "affectionate" were not found to be undesirable for either sex using any scale, a significant difference was found in degree of desirability ratings for men and for women when Bem's scale was used. This supported the first of Bem's hypotheses: that masculine and feminine characteristics are not polar opposites, but rather are characteristics which are perceived as more socially desirable for one sex than for the other. Bem's description of items as "sex-appropriate" and "sex-inappropriate" (1974, p. 158) is inappropriate. Most items are merely more or less appropriate for each sex; the importance of this distinction can be seen in the context of Stricker's (1977) analysis of the Broverman (1968, 1970, 1972) studies.

Some BSRI items may violate Bem's assumption of relative desirability. Edwards and Ashworth's judges rated "masculine" and "feminine" as undesirable (below 4.0) for both a female and a male. The Feminine items "gullible" and "childlike" were also rated undesirable for both a male and a female (Edwards & Ashworth, personal communication). Although the judges in the present study did not have the opportunity to make ratings of undesirability, both males and females did give ratings of these four items close to "not at all desirable" (below 2.0) in the same pattern as Edwards and Ashworth's judges. Thus, these items appear not only relatively more desirable for either a man or a woman, but also relatively undesirable for one or both sexes.

Waters, Waters, and Pincus's (1977) factor analysis of BSRI Masculine and Feminine item self-report responses found that "masculine" and "feminine" loaded on a factor associated with subjects' gender and that "gullible" and "childlike" loaded negatively on a masculine sex-type factor. The present analysis leads to a tentative concurrence with their suggestion that

these items should be deleted to increase the homogeniety and interpretability of the *M* and *F* scales. However, deletion of "gullible" and "childlike" would increase the social desirability of the *F* scale relative to the *M* scale. The resulting imbalance should increase the correlation of Femininity and Androgyny scores with social desirability. Before deleting items, investigators should ascertain the impact of deletion on the problem of social desirability response sets.

With the exception of these items, the results suggest that the difference between Bem's (1974) and Edwards and Ashworth's (1977) results was attributable to the difference in the rating instructions and anchored rating scale and, with respect to masculine and feminine BSRI items, not to the other hypotheses raised by Edwards and Ashworth.

Bem found no significant differences between the ratings of male and female judges on the 20 BSRI neutral items. This finding was replicated for 10 of the BSRI neutral items. For the other 10 neutral items identified in the results section, one or both sexes judged the items as more desirable "for a woman" or "for a man" in American society. In contrast to Bem, these items were not rated as equally desirable in this study. Bem's failure to find a significant difference between "for a man" and "for a woman" may be attributable to reduced statistical power, since only 10 judges rated the positive neutral items and 15 judges rated the negative neutral items. Furthermore, Bem's judges were rating these items together with a much larger sample of potentially neutral items (Bem, personal communication). Ratings of neutral BSRI items are highly skewed, with means at the lower or upper ends of the rating scale. As long as the Neutral scale score is not used (Bem, 1977), the issue is moot. However, it should no longer be assumed that Bem's neutral items are equally desirable in American society for a man and for a woman. This may open to reinterpretation the near zero correlations between the Neutral scale and Androgyny scores reported by Bem (1974). It is important to note that the BSRI was designed for the purpose of assigning subjects to relative categories for social psychology experiments testing Bem's hypotheses. The temptation to apply the BSRI scales in other contexts (to which several have succumbed) should be tempered by reexamination of the methods of construction, scoring, reliability estimation, and validation employed to date.

Conclusion

Prompted by Edwards and Ashworth (1977), Bem's (1974) BSRI item selection strategy was reexamined, using her anchored rating scale and instructions. The results replicated Bem's for 18 masculine and 19 feminine items and suggest that the failure to replicate reported by Edwards and Ashworth resulted from differences in the anchored rating scale and instructions.

One masculine and three feminine items were found to violate Bem's assumption of relative desirability; thus, they should be deleted, given minimal impact on the correlation between social desirability and BSRI scores. Since half the neutral items were rated significantly more desirable for a man or for a woman in this study, it is recommended that it no longer be assumed that neutral items are neutral with regard to sex.

References

Bem, S. L. The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 1974, *42*, 155–162.

Bem, S. L. Sex-role adaptability: one consequence of psychological androgyny. *Journal of Personality and Social Psychology*, 1975, *31*, 634–643.

Bem, S. L. Probing the promise of androgyny. In A. G. Kaplan & J. P. Bean (Eds.), *Beyond Sex Role Stereotypes*. Little, Brown, and Company, 1976, 48–62.

Bem, S. L. On the utility of alternative procedures for assessing psychological androgyny. *Journal of Consulting and Clinical Psychology*, in press.

Bem, S. L., & Lenney, E. Sex typing and the avoidance of cross-sex behavior. *Journal of Personality and Social Psychology*, 1976, *33*, 48–54.

Bem, S. L., Martyna, W., & Watson, C. Sex-typing and androgyny: further explorations of the expres-

sive domain. *Journal of Personality and Social Psychology*, 1976, *34*, 1016–1023.

Broverman, I. K., Broverman, D. M., Clarkson, F. E., Rosencrantz, P. S., & Vogel, S. R. Sex-role stereotypes and clinical judgments of mental health. *Journal of Consulting and Clinical Psychology*, 1970, *34*, 1–7.

Broverman, I. K., Vogel, S. R., Broverman, D. M., Clarkson, F. E., & Rosencrantz, P. S. Sex-role stereotypes: A current appraisal, *Journal of Social Issues*, 1972, *28*, 59–78.

Edwards, A. L., & Ashworth, C. D. A cross validation of item selection for the Bem Sex Role Inventory. *Applied Psychological Measurement*, 1977, *1*, 501–507.

Kelly, J. A., & Worell, L. Parent behaviors related to masculine, feminine, and androgynous sex role orientations. *Journal of Consulting and Clinical Psychology*, 1976, *44*, 843–851.

Rosencrantz, P. S., Vogel, S. R., Bee, H., Broverman, I. K., & Broverman, D. M. Sex-role stereotypes and self-concepts in college students, *Journal of Consulting and Clinical Psychology*, 1968, *32*, 287–295.

Ruble, D. N., Croke, J. A., Frieze, I., & Parsons, J. E. A field study of sex-role attitude change in college women, *Journal of Applied Social Psychology*, 1975, *5*, 110–117.

Spence, J. T., Helmreich, R., & Stapp, J. Rating of self and peers on sex role attributes and their relation to self-esteem and conceptions of masculinity and femininity. *Journal of Personality and Social Psychology*, 1975, *32*, 29–39.

Strahan, R. F. Remarks on Bem's measurement of psychological androgyny: alternative methods and a supplementary analysis. *Journal of Consulting and Clinical Psychology*, 1975, *43*, 568–571.

Striker, G. Implications of research for psychotherapeutic treatment of women. *American Psychologist*, 1977, *3*, 14–22.

Wakefield, J. A. Jr., Sasek, J., Friedman, A. F., & Bowden, J. D. Androgyny and other measures of masculinity-femininity. *Journal of Consulting and Clinical Psychology*, 1976, *44*, 766–770.

Waters, C. W., Waters, L. K., & Pincus, S. Factor analysis of masculine and feminine sex-typed items from the BSRI. *Psychological Reports*, 1977, *40*, 457–570.

Author's Address

Robert D. Abbott, Educational Psychology, DQ-12, University of Washington, Seattle, WA 98195