The Journal of Genetic Psychology, 1977, **130**, 117-120.

REPLY TO A. R. JENSEN'S COMMENTS ON M. W. SMITH'S ANALYSIS OF CULTURE BIAS IN THE STANFORD-BINET INTELLIGENCE SCALE* [1]

University of California, Berkeley

MARGOT WIESINGER SMITH

SUMMARY

This article is the author's reply to A. R. Jensen's critique in the present issue on her paper, "Alfred Binet's remarkable questions: A cross-national and cross-temporal analysis of the cultural biases built into the Stanford-Binet Intelligence Scale and other Binet tests," which appeared in *Genetic Psychology Monographs*, 1974, **89**, 307-334.

Jensen comments that a conclusion in my research on the Stanford-Binet Intelligence Scale is "based on serious factual errors and is therefore wholly unfounded." Jensen's comments are best answered point by point.

1. To quote, "Smith's evidence for temporal change in the difficulty level of Stanford-Binet (S-B) items is given on Tables 3 and 4 of her article . . ." (3, pp. 113-115). In fact, the findings of percent-passing scores and mean *IQ* decreases over time were derived from several data sources. These findings included (*a*) the need to revise the S-B every 20 years because of test obsolescence due to culture change; (*b*) the data from cross-national comparisons shown on Table 4, where in every case tests later in time had lower percent-passing scores; (*c*) P. E. Vernon's data on London children, cited; (*d*) Scottish data on tests given to normative samples of 11-year-olds in 1932 and 1937 (4, pp. 329-330). Additional evidence that scores are lower for groups with the same characteristics as the original normative sample, only tested at

* Received in the Editorial Office, Provincetown, Massachusetts, on June 4, 1975.

[1] It is the policy of this journal to publish a critique of an article published here or in *Genetic Psychology Monographs*, with the understanding that the author of the original article has an opportunity to reply. However we do not publish prolonged controversial series or replies, and readers may draw their own conclusions from the material presented. Readers who have further questions should write directly to the authors.

a later moment in time, can be examined on Table 4 (4, p. 331) by comparing figures from the 1916 and 1937 Stanford-Binet's: 1916, 66%; 1937, 62%.

2. Jensen is correct in stating that the 1937 and 1960 percent-passing scores in Appendix B of the Terman/Merrill Manual (7) are not comparable, and I appreciate his effort in writing to Dr. Merrill to verify this fact. It is unfortunate that the Terman/Merrill table was "insufficiently labeled" (2), as I assumed that the percent-passing scores followed the conventions of the other 11 Binet revisions in the analysis and were based on *CA*. There was no indication in the text or on the table in Appendix B that the figures were not comparable. However, the use of *MA* in constructing the 1960 Stanford-Binet has profound theoretical implications which go beyond my original article. As Terman and Merrill said in 1937 (6, p. 10) "increase in percent-passing by mental age is better (as a measure of validity for selection or rejection of test items) but *exclusive reliance* on this technique predetermines that the scale based upon this criterion *will measure approximately the same functions* as that used in selecting the mental age group" (emphasis mine). Terman and Merrill imply that a tautological relationship between test and test score would result from this procedure. Merrill herself states that "the difficulty of items (in the 1960 test) was determined by computing the percentage of children passing at successive mental ages. This procedure predetermines that a scale based upon this criterion will measure the same functions that were used in selecting the mental age group, in this case, mental age determined by the scale" (7, pp. 23-24). Historically, the basic data for construction of Binet scales have been the percent of children of a certain chronological age who answer a question correctly. In the 1960 scale, the basic data for determining the level of difficulty for Binet questions were the percent of children of a certain mental age, or test score, who answered a question correctly. The final test score (*IQ*), then, has a circular relationship to the test itself, as it was constructed with use of a sample of children defined as homogeneous by test score, or mental age, not by the objective criterion of chronological age.

How would the construction of the 1960 test using mental age affect the norms? Merrill noted that *MA* percent-passing curves were more sharply inflected (7, p. 29). Garfinkel and Thorndike commented that "ninety-eight of the 122 items appear to have become more age specific over the 40 year period" (1, p. 4). This seems odd, as the 1960 and 1972 samples were not as age-specific as to chronological age as the 1937 and other samples. How the use of *MA* has affected test scores since 1960 can only be determined by examining data, perhaps of longitudinal studies. Researchers since 1960 may

be able to tell whether scores have become more randomized, more polarized, or vary in other ways from previous patterns.

The fact that the data from the 1937 and 1960 Stanford-Binet scales are not comparable does not alter my finding that test scores are lower for groups with the same characteristics as the original normative sample when tested at a later moment in time, as this finding was based on an array of data from several sources, not just on one data point. Jensen's scholarship is appreciated. In his pointing up the unique construction of the 1960 Stanford-Binet Intelligence Scale, problems with *IQ* scores and distortions in the test can be examined. For further discussion of the matter, please refer to my article, "Distortions in the 1972 norms of the Stanford-Binet Intelligence Scale" (5).

3. Jensen comments that the error of comparing 1937 and 1960 data is repeated in Figures 3, 4, and 5 of my paper. However, the 1960 data are just one point among many percent-passing curves. As Dr. Merrill says, "No single figure gives an adequate basis for assessing the difficulty of the subtests of a scale (7, p. 31). Unfortunately, she never released the percent-passing data for the 1960 tests, either *CA* or *MA*.

4. As Jensen noted, the data for black children were not adequate, as they were given by grade and not by age, and the sample was not selected according to the same criteria as other Binet revisions. I commented on this twice, once on page 318 and again on Table 2. That makes the point twice in the article, which ought to be enough.

5. The following is in response to the final paragraph of his comments that "serious doubts seem justified concerning an analysis of cultural biases in Binet items that reports quantitative results based on a handpicked selection of only eight items . . ." In the body of the article (4, pp. 320-326) 19 test items are named, and I say "Most items have a dispersion of percent-passing curves in the middle range *illustrated* in Figures 2, 3, 4, and 5 . . ." (4, p. 320). Figures 1, 6, 7, and 8 are used to illustrate an empirical means of determining the cultural loading of Binet test items. As Jensen said, there are 192 items on the Stanford-Binet; however, many of these are duplications and other versions of the same questions. Binet's original questions numbered 65, counting both the 1908 and 1911 versions. If anyone would like to see percent-passing curves for the other Binet questions, I would be glad to oblige. The eight figures in the article were representative, not the total data available for analysis.

I reiterate and reaffirm my original findings: (*a*) Binet questions have cultural loadings which can be empirically measured. (*b*) Each Binet ques-

tion has a cultural loading, varying in degree from item to item. (*c*) The cultural loading of a test item binds the test to a specific time and locale, which has been adjusted for over the years by periodically revising the test to omit the obsolete and the exotic. (*d*) Binet test scores are consistently lower for normal populations with the same characteristics as the normative sample, only tested at a later moment in time. Binet tests are carefully and rationally designed instruments with certain properties which operate independent of the characteristics of the sample.

REFERENCES

1. GARFINKEL, R., & THORNDIKE, R. Binet item difficulty then and now. Unpublished manuscript, Teacher's College, Columbia University, New York, 1974.
2. JENSEN, A. R. Personal communication, 1975.
3. ———. An unfounded conclusion in M. W. Smith's analysis of culture bias in the Stanford-Binet Intelligence Scale. *J. Genet. Psychol.*, 1977, **130,** 113-115.
4. SMITH, M. W. Alfred Binet's remarkable questions: A cross-national and cross-temporal analysis of the cultural biases built into the Stanford-Binet Intelligence Scale and other Binet tests. *Genet. Psychol. Monog.*, 1974, **89,** 307-334.
5. ———. Distortions in the 1972 norms of the Stanford-Binet Intelligence Scale. Paper presented at American Psychological Association Meetings, Chicago, Illinois, August, 1975.
6. TERMAN, L., & MERRILL, M. Measuring Intelligence. Boston: Houghton Mifflin, 1937.
7. ———. Stanford-Binet Intelligence Scale. Manual for the Third Revision, Form L-M. Boston: Houghton Mifflin, 1960.

900 San Benito Road
Berkeley, California 94707