Psychology and Aging
2004, Vol. 19, No. 3, 379–393

0882-7974/04/$12.00 DOI: 10.1037/0882-7974.19.3.379

Category Norms as a Function of Culture and Age: Comparisons of Item Responses to 105 Categories by American and Chinese Adults

Carolyn Yoon and Fred Feinberg
University of Michigan

Ping Hu
Chinese Academy of Sciences

Angela Hall Gutchess, Trey Hedden, and Hiu-Ying Mary Chen
University of Michigan

Qicheng Jing and Yao Cui
Chinese Academy of Sciences

Denise C. Park
University of Illinois at Urbana–Champaign

Understanding how aging influences cognition across different cultures has been hindered by a lack of standardized, cross-referenced verbal stimuli. This study introduces a database of such item-level stimuli for both younger and older adults, in China and the United States, and makes 3 distinct contributions. First, the authors specify which item categories generalize across age and/or cultural groups, rigorously quantifying differences among them. Second, they introduce novel, powerful methods to measure between-group differences in freely generated ranked data, the rank-ordered logit model and Hellinger Affinity. Finally, a broad archive of tested, cross-linguistic stimuli is now freely available to researchers: data, similarity measures, and all stimulus materials for 105 categories and 4 culture-by-age groups, comprising over 10,000 fully translated unique item responses.

The past decade has seen a marked increase in cross-cultural research in both cognitive and social psychology. Differences in cognitive and reasoning processes prevalent in Eastern and Western cultures have emerged as a topic of particular interest (e.g., Bond, 1991; Chan, 1996; Fiske, Kitayama, Markus, & Nisbett, 1998). More recently, researchers have emphasized how investigations into the interplay between aging and cultural differences in cognitive processes—particularly so between East Asian and Western cultures—can help distinguish the aspects of cognitive aging that are culturally invariant from those that depend on culture (Park, Nisbett, & Hedden, 1999). Research on cross-cultural and age differences in cognition, however, has been hampered by the lack of appropriate normed verbal materials.

Given the time-consuming and costly nature of conducting norming studies, it is unsurprising that the availability of norms is limited. With respect to category norms comparing younger and older adult groups within a culture, the limited availability is further attributable to a general assumption among researchers that semantic categories remain relatively unaffected by normal aging, that is, that the structure of semantic categories does not qualitatively vary as individuals age. That semantic knowledge remains constant across the lifespan appears to be inferred primarily from two sources. First, researchers have documented a lack of age-related declines in performance on standard vocabulary tests (e.g., Salthouse, 1993); in fact, many studies have reported improvement in vocabulary scores with age (see Kausler, 1991, for a review). Second, notwithstanding the slower responses typically observed with aging (e.g., Birren & Fisher, 1995), results of semantic priming studies indicate that facilitative priming effects do not degrade across the lifespan (e.g., Balota, Black, & Cheney, 1992; Light & Singh, 1987).

On the one hand, it stands to reason that knowledge representations of common categories (e.g., type of fruit) remain intact if people, as they age, continue to be functional in everyday settings. On the other hand, it is not clear whether this stability of representations holds for categories that draw less heavily on collectively shared knowledge (e.g., type of herbal medicine or food flavoring). Thus, one of the main goals of the present study was to investigate the extent to which this widely held view, that categories remain largely invariant with age, does indeed hold true. We did so by assessing the degree to which normed responses remain constant for certain types of categories (and change for other

Carolyn Yoon and Fred Feinberg, University of Michigan Business School, University of Michigan; Ping Hu, Qicheng Jing, and Yao Cui, Institute of Psychology, Chinese Academy of Sciences, Beijing, China; Angela Hall Gutchess, Trey Hedden, and Hiu-Ying Mary Chen, Department of Psychology, University of Michigan; Denise C. Park, Department of Psychology, University of Illinois at Urbana–Champaign.

This study was supported by Grant R01 AG015047 from the National Institute on Aging. We thank the following people for their input at various phases of the project: Phaythoune Chothmounethinh, Daphne Fuad, Gang Guo, Suzanne Hambright, Alice Huang, Paula Johanes, I-Lin Kuo, Berinda Lee, Michelle Lee, Jason Pau, Jiaxuan Shen, Twila Tardiff, Mary Wagner, Michael Wagner, Vickie Yeung, Hong Yuan, and Zhen Zeng. We extend special thanks to Wei Huang and Yanfen Zhang for their assistance and expertise.

To retrieve the broad archive of tested, cross-linguistic stimuli, go to http://agingmind.cns.uiuc.edu/Cat_Norms/

Correspondence concerning this article should be addressed to Carolyn Yoon, University of Michigan Business School, 701 Tappan Street, Ann Arbor, MI 48109-1234. E-mail: yoonc@umich.edu

types) across age groups within the United States and in mainland China.

The second goal of the present project was to explore how age differences in category norms *vary* across cultures. There may, for example, be fairly common categories that remain stable for age groups not only within a culture but across cultures as well. We therefore compared category norms for both younger and older adults in China with their age-matched counterparts in the United States; we further sought to identify metatypes of categories that generalize across age groups not only within a culture but also across cultural groups.

As we discuss at length, there are considerable problems associated with analyzing the type of responses collected in this area of research—freely generated rank-ordered data—and additional ones that arise in comparing results across participant groups. Specifically, the data are nominal, are sparse, and obey order restrictions, calling for methods beyond standard correlational analyses. Our second goal was therefore methodological: to present rigorous measurement techniques for such data. The methodology developed, combining the rank-ordered logit model and Hellinger Affinity, can be readily deployed in a wide variety of studies and offers several distinct advantages over Pearson- and Spearman-based measures.

Conducting cross-cultural research with East Asian and Western subjects is a particularly daunting undertaking, owing to the need to prepare and verify verbal stimuli in highly disparate languages, with unrelated, complex orthographic systems. A final goal of this research was to create an archive of reliably translated item responses for 105 categories. To our knowledge, the project Web site offers the most extensive repository of cross-linguistic categorical item names presently available: over 10,000 unique items, fully translated, including many absent from standard reference works, such as those dealing with popular culture or regional custom.

Prior Norming Studies and Cross-Age and Cross-Culture Comparison

Researchers wishing to understand the role of such important metavariables as age and culture in "universal" theories of cognition and memory find their task complicated by the lack of confluence between stimulus materials across studies. Creators of several independent sets of U.S.-based category norms have attempted to remedy this situation, among them Battig and Montague (1969), Cohen, Bousfield, and Whitmarsh (1957), McEvoy and Nelson (1982), and Shapiro and Palermo (1970). However, each of these studies was conducted exclusively with younger individuals. Over the intervening decades, researchers in a number of countries have attempted to validate norms for various age or cultural groups using different subsets of the 56 Battig and Montague (1969) categories: for example, Australia (Casey & Heath, 1988), Britain (Hampton & Gardiner, 1983), Ireland and Scotland (Brown & Davies, 1976), Israel (Henik & Kaplan, 1988–1989), New Zealand (Marshall & Parr, 1996), Portugal (Pinto, 1992), and Spain (Pascaul & Musitu, 1980). Relatively few researchers (e.g., Henik & Kaplan, 1988–1989) have compared their derived norms with those of Battig and Montague for Americans. A related line of research has provided norms for small numbers of categories on preschool and school children in various cultures, from countries such as Germany (Hupbach & Mecklenbräuker, 1998), Spain (Goikoetxea, 2000), and the United States (Posnansky, 1978).

Comparing Norms Across Cultures and Age Groups

The single published study to date to have compared category norms for younger and older Americans is by Howard (1980), who tested a subset of 21 of the Battig and Montague (1969) category norms to determine whether they were appropriate for use with both age groups. Since then, only one other category-norming study for different age groups within a culture has been conducted. Boccardi and Cappa (1997) collected category norms for younger and older groups in Italy and drew conclusions similar to Howard's about the stability of category responses across the lifespan. To our knowledge, only two extant studies have collected category-norming data for age-matched samples across different cultures. Brown and Davies (1976) collected norms for 15 categories on university students in Scotland and Northern Ireland and reported significant differences across culture. Hasselhorn, Jaspers, and Hernando (1990) obtained norms for 10 categories from German children and compared the results with norms for German adults and with Posnansky's (1978) norms for American children; they also found differences across culture, in addition to differences across age for their German sample.

Eastern and Western Norms

Although prior evidence is admittedly limited, it nonetheless suggests substantial differences across age and culture even when comparisons are between Western cultural groups. Given the far greater linguistic, historical, political, and societal differences between East Asian and Western cultures, one might expect any such disparities in category norms to be magnified. Indeed, a limited number of cross-cultural studies on judgments of category prototypes seem to suggest greater cross-cultural variations in category structure as cultures are more distinct from each other, for example, Taiwanese–Chinese and American (Lin & Schwanenflugel, 1995; Lin, Schwanenflugel, & Wisenbaker, 1990). Given the practical difficulties of conducting rigorous comparisons between even closely related groups—a topic we address at length—it is unsurprising that virtually all broadly available category-norming data were obtained primarily on Western samples. The sole exception is the Jeng, Lai, and Liu (1973) study, which reported norms in Chinese and English for 54 of the Battig and Montague (1969) categories in Taiwan.[1]

Method

We wish to present a full account of a relatively lengthy chain of procedures, starting with the collection of individual participants' data in two age groups and languages, and successively processing these data to

[1] The generalizability of Jeng et al.'s (1973) normative data is limited in several ways: All participants had taken at least 6 years of English classes in Taiwan; responses by Taiwanese students may tend to reflect items in Chinese that are most readily translated into English; the rank order of frequency data is not reported, making post hoc inference difficult; and, finally, responses generated by their Taiwanese participants differ substantially from those of the mainland Chinese participants in the present study.

Table 1
Age, Education, Health, and Gender Characteristics for Americans and Chinese, by Age Group

Culture and age group	*n*	Age (years)		Education (years)		Self-rated health status		Gender	
		M	*SD*	*M*	*SD*	*M*	*SD*	Male	Female
American									
Young	113	19.83_{a}	1.66	13.92_{a}	0.12	3.59_{a}	0.82	39%	61%
Old	103	66.85	4.05	15.67_{b}	0.14	3.84_{a}	0.81	41%	59%
Chinese									
Young	100	20.09_{a}	1.04	14.04_{a}	0.75	$3.31_{a,b}$	0.73	39%	61%
Old	100	64.68	3.38	16.73_{b}	1.35	3.24_{b}	0.67	42%	58%

Note. Health status was assessed on a 5-point scale (1 = *much worse than average*, 2 = *worse than average*, 3 = *average*, 4 = *better than average*, 5 = *much better than average*). Means in the same column that do not share subscripts differ at $p < .05$.

the point where defensible inferences can be made. At heart, this involves quantification and processing of an enormous body of purely verbal, individual-level data in two strikingly disparate languages. Therefore, unlike in most prior research, we must discuss the problems arising in translating thousands of items in over 100 categories, and also in dealing statistically with a daunting data type: sparse, rank-ordered, individual-level nominal data. We believe this is important not only for interpretation but for replicability both within and outside the groups in the present study. What follows, then, is a line of reasoning leading from choice of participants and stimulus materials, through data collection; an unusually complex process of coding, categorizing, and cross-translation; the application of two analysis methods novel in this area; and, in culmination, inferences about categories, age groups, and cultures.

Participants

One hundred thirteen younger adults (17–23 years) from the University of Michigan and 103 community-dwelling older adults (60–75 years), constituting the American cultural group, were tested in Ann Arbor, Michigan. The younger participants were recruited from the subject pool in the Psychology Department at the University of Michigan. Older participants were community-dwelling adults recruited from a senior community center located in Ann Arbor. One hundred younger Chinese students (18–23 years) recruited from three universities (Beijing Normal University, Capital Normal University, and Aeronautics and Space University) and 100 older Chinese individuals (59–76 years) from the local community were tested at the Institute of Psychology, Chinese Academy of Sciences, in Beijing, China. Summary age statistics for all groups appear in Table 1.

Participants from all four samples were well educated, with mean (and median) levels of at least 2 years of college education. All participants rated their general health as being at least average (greater than 3 on a 5-point scale). All younger study participants were college students. Eighty percent of the older American adults and 91% of the older Chinese adults were retirees, and these groups had similar median numbers of years since retirement (*Mdn* = 7.0 and 6.0 years for Americans and Chinese, respectively). The remaining older adults reported working either part time or full time in various professional occupations. Finally, the gender distribution across the four culture-by-age groups was approximately equivalent. Whereas great care was used to ensure that the samples were demographically *similar,* they were deliberately not forced to be demographically *representative:* Given the much greater differences in the urban and rural populations in China relative to their counterparts in the United States, *within*-culture demographic representativeness would have made *cross*-cultural inferences impossible.

Stimulus Materials and Choice of Categories

The practical difficulties of developing and comparing category norms across different cultures are magnified when the cultures being compared are as dissimilar as the American and Chinese in terms of socialization, experiences, and, in particular, language. In choosing categories, we chose to include those developed by Battig and Montague (1969) because they have been the most widely used among published norms. Of their 56 categories, 9 were excluded because their overt local reference rendered them less suitable for cross-cultural comparisons (e.g., elective office, kind of money, member of the clergy).

We considered it important to include categories that were likely to vary in terms of how many different item responses would be generated by participants, within or across culture.[2] This formed the basis for our selection of the remaining 58 categories, which comprised 46 categories taken from McEvoy and Nelson (1982) and 12 additional categories that originate here. For example, "farm animal" (from McEvoy & Nelson, 1982) would be expected to yield relatively few instances in both cultures, because it constitutes a rather limited, delineated category. By contrast, "mythical being" may be an intrinsically more differentiated category, likely to generate a greater number of item responses. We also included categories that would likely display greater differentiation in one culture than the other. Interesting examples are provided by two types of relations included in the study: familial and nonfamilial. Because of a greater emphasis in the Chinese culture on kin relationships, "relatives" was expected to be a more differentiated category for the Chinese than for the Americans (and proved to be dramatically so in the actual data, with 37 unique terms used by the Americans and 124 by the Chinese). On the other hand, Americans are more likely than Chinese to place emphasis on professional or other relationships, so that "nonrelative relations" was expected to be relatively more differentiated for them (again borne out in the data, with 185 distinct items for the Americans and 131 for the Chinese).

We further aimed to construct a relatively heterogeneous set of categories by selecting partonomies (e.g., part of the human body; part of a boat) as well as categories with a variety of taxonomic levels (Rosch, 1973, 1978). For example, the category "fish" is at the "basic" level of specificity, the most natural and preferred level at which individuals conceptually

[2] Throughout, the term *item* is reserved for the objects contained within categories; *categories* always denotes collections of items or objects; and *group* refers to one of two cultural groups (Chinese or American) or age groups (younger or older) or one of the four possible combinations of culture by age in the study.

carve up their experience. By contrast, "tropical fish" is at a relatively subordinate level, whereas "animal (wild)" is superordinate. We note in passing that some categories, such as "fish" or "fruit," may be basic for some respondents but not for others, even within a culture-by-age group; one must take care not to interpret the categories as if all members in a particular group view them in an identical hierarchical structure, and we do not do so here. It is important to realize that concepts such as subordinate and superordinate cannot themselves be presumed culturally invariant: A category such as "type of relative" may, for example, be perceived as more of a vague, catch-all category in China, where the language itself is replete with highly specific familial distinctions—as in the sharp divide between maternal and paternal kinship—which a Westerner would hardly think to articulate (Huang & Jia, 2000, provide explicit comparisons).

Thus, 108 categories were selected for testing. All 108 were translated into Chinese separately by two research assistants fluent in both languages, along with a translation panel consisting of psycholinguists, professional translators, and a group of psychologists with native command of both English and Mandarin Chinese.[3] Despite these steps taken to ensure translation accuracy, three category names (social activity, form of government, and branch of armed services) appeared to have been misinterpreted by Chinese adult participants owing to subtle problems in translation, and were thus excluded from further analysis. The final list of 105 categories appears in the Appendix.

Procedures

Experimental sessions were conducted in groups of 5 to 25 participants each. Participants were given general instructions, followed by several practice examples. Chinese participants were given both verbal instructions in Mandarin and written instructions in simplified Chinese characters; Americans were provided English equivalents. Participants were then given four booklets, in counterbalanced order, each containing 27 pages with a category name printed at the top; category order was randomized within booklets. On each page, the category name was followed by five blank lines on which participants were asked to list five items belonging to that category, in the order in which they came to mind. For the category "reading material" only, participants were also given additional written instructions that they not write down names of specific titles.

Younger participants in both cultural groups were given 20 min to complete the entire booklet, whereas older adults were allowed 30 min; informal observation suggested time was not a factor in ability to successfully complete the task. Upon completion of the first booklet, three additional booklets were administered in the same manner, so that each participant's data consisted of four separate paper booklets, comprising a total of 108 categories (from which 3 were later excluded), each with five item responses.

Following the main task, two short questionnaires were administered, one eliciting demographic information and the other relating to health status. Participants were then debriefed, thanked for their participation, and dismissed. Younger Americans received credit in the form of partial fulfillment of their subject pool requirements in a psychology course; the other groups received monetary compensation for their participation. The entire session lasted approximately 90 min for younger adults and 140 min for older adults.

Results

Translation of item responses between Chinese and English posed challenges well beyond those encountered in most former studies of categorization. Some of these descended from the pictographic nature of written Chinese, others from typical verbal usage patterns, and still others from Chinese's uncommon morphology (i.e., the way in which new, compound words are created). English is notorious for its weak morphology, even among Germanic languages (Avgerinos & Theofilos, 1995; Piotrowski, 1998). Perfetti, Zhang, and Berent (1992) and Huang and Hanley (1994) present excellent overviews of how Chinese and English differ from both a phonological and a morphological perspective. We discuss the most common problems of translation and "rendering"—both within each language and between them—to convey how item categorizations were determined.[4]

Identicality Within a Language

Within both languages, it was necessary to decide when two item names were merely different labels for an identical, underlying thing. One might state this problem in terms of whether item codings are primarily lexical or primarily semantic. With few exceptions, we hewed to the former. That is, except for very obvious variants, misspellings, or abbreviations, we treated item names like *couch* and *sofa* as distinct. Although a more semantic coding scheme would also have been possible, it would have required a large degree of inference about participants' intentions and meanings—must participants always mean the same thing by *couch* and *sofa* or by *fridge* and *refrigerator?*—a process fraught with subjectivity, particularly so across different languages. When item names were obviously identical or nearly so (e.g., *high-rise* and *hi-rise* for "human dwelling"), they were counted as a single item and assigned a consistent item number valid across age and culture groups. All item conjoinment was decided in tandem with the translation panel, with the admittedly high standard that any such item pairs would be considered definitionally or trivially identical by all native speakers.

Colloquial Usage and Pictography in Chinese

In Chinese, certain terms are deemed identical in their colloquial usage (e.g., 大象 and 象 [elephant]; 老虎 and 虎 [tiger]), for which there is no analogous convention in English. All such responses were considered individually by the translation panel in consultation with a Chinese linguist[5] and coded, where appropriate, as identical responses. In such cases, items were considered identical only if any native speaker would consider them so.

Many-to-One Item Mappings Between Languages

Another problem, and generally a more difficult one to rectify, involves how finely each language distinguishes a set of items, or even what appears to be a *single* item. An excellent example is provided by a precious stone mentioned frequently by both the American and Chinese samples: jade. Americans were unanimous in calling a jade a *jade;* English has one collective term that subsumes all forms of the stone. In contrast, Chinese participants

---

[3] Translation results were further checked by those among the authors fluent in both languages. We also wish to thank Yanfen Zhang, Gang Guo, Wei Huang, Jiaxuan Shen, and Twila Tardiff for lending their expertise.

[4] All item correspondences, for each of the 105 categories, are detailed on the project Web site, http://agingmind.beckman.uiuc.edu/Cat_Norms, or at its mirror, http://agingmind.cns.uiuc.edu/Cat_Norms/.

[5] We thank Jiaxuan Shen, director of the Institute of Linguistics, Chinese Academy of Social Science, Beijing, China, for lending his linguistic expertise.

consistently made reference to at least four terms that correspond to the English *jade:* 碧玉 (good quality with a bluish-green tint); 翠, 翡翠 (often referred to as *jadeite,* highly valued for jewelry and sculptures because of its brilliant emerald color); 玉 (valuable but slightly softer than jadeite and often veined; used in bowls and vases and also as a general term for Chinese jade used in objects); and 玉石 (unprocessed, with emphasis on the stone or boulder form). To Chinese speakers, these are all readily distinguished, and it is unclear the extent to which they are conceived as different *forms* of the same underlying substance, "jade" (much as English speakers generally do not conceive of ruby and sapphire as different forms of the mineral corundum, so that they are intrinsically more similar than either is to emerald, a form of beryl). It would be incorrect to infer that Chinese is uniformly more "finely grained," as English provided similar examples: *stout, lager, ale,* and *beer* correspond to the Chinese 啤酒 and 干啤, 干啤酒, and there are substantially finer distinctions among certain foods, such as breads, as well as common consumer goods.

One-to-Zero Item Mappings Between Languages

Even when certain words or terms were perfectly unambiguous in Chinese, even fully bilingual speakers were hard pressed for suitable English equivalents. For the category "mythical being," a relatively common response was 猪八戒, which translates into *the pig monk,* a concept incomprehensible to most Westerners. Other common mythical beings were the *xiao tian dog* (杨戬和哮天权) and the *goddess who mended the sky* (女娲). As one might imagine, none of these was mentioned by the American participants. Of course, the reverse occurred as well, as Chinese speakers were unfamiliar with such quintessentially Western items and concepts as *coyote* (wild animal), *Lego* (toy), and *cyclops* (mythical being). Given 105 categories, four culture-by-age groups, and several hundred responses on average for each, accurate identification of nonoverlapping categories was a daunting, yet crucial, task. Even with tens of thousands of unique individual item responses, only a very small number could not be rendered faithfully from one language into the other, and in all such cases, this indicated that the item in question was unique to one cultural group.

Preparation of Individual-Level Data

For each culture-by-age group, data were compiled at the individual level for each of the 105 categories. All responses by Chinese participants were rendered separately into English by two bilingual translators; any remaining differences were resolved via discussion among them, the translation panel, and the authors. The rank position of each item in a participant's response sequence for a given category was also recorded (and figures prominently in the forthcoming analysis).

To ensure consistency and reasonableness of the coding scheme across both American and Chinese cultures, we established a firm set of guidelines. First, all legible responses were recorded, with any misspellings (e.g., homonyms) corrected. Second, when a participant included two or more words on a single line, only the first was retained (e.g., *bucket* for *bucket, pail*). Third, any superfluous articles accompanying a word were discarded (e.g., *the, a, an*), as were pluralizations. Fourth, any nontrivial elaborations (e.g., *small button* and *button*) were counted as distinct items, consistent with codings in prior studies (e.g., Battig & Montague, 1969). Finally, for Chinese responses only, an additional criterion was imposed to accommodate conventions specific to the Chinese language. As mentioned earlier, in Chinese, certain terms are deemed identical in their colloquial usage (e.g., 大象 and 象 [elephant]), for which there are no analogous examples in English. All such responses were considered individually and discussed extensively to determine how they should be coded; on the basis of these discussions, item terms were merged when appropriate. In all lists and tables containing the data, Chinese terms merged in this manner have been retained and are specifically identified as such (e.g., 象, 大象).

For each category, a master list of numerical codes was created to facilitate response comparisons across the culture-by-age groups. In all project files, items are labeled with these category-specific numerical codes, as well as numbers for each category (1–105). Each participant's response data within a category therefore consists of five category-specific item numbers, in order. A schematic of the individual-level data, and how they are analyzed, appears in Figure 1. For Category 68, "musical instruments," the (ordered) responses of Subject 1001 were {saxophone, trumpet, tuba, flute, clarinet}, which translate to {40, 46, 47, 21, 12}, as shown in the figure. The data, for this example, continue for all 113 younger American and 103 older American participants. As discussed at length below, these raw item codes form the "inputs" for our analysis: They are fed into a model that reranks them, after which an "affinity" or comparison score is calculated for the two culture-by-age groups (younger vs. older Americans) in that category (musical instruments).

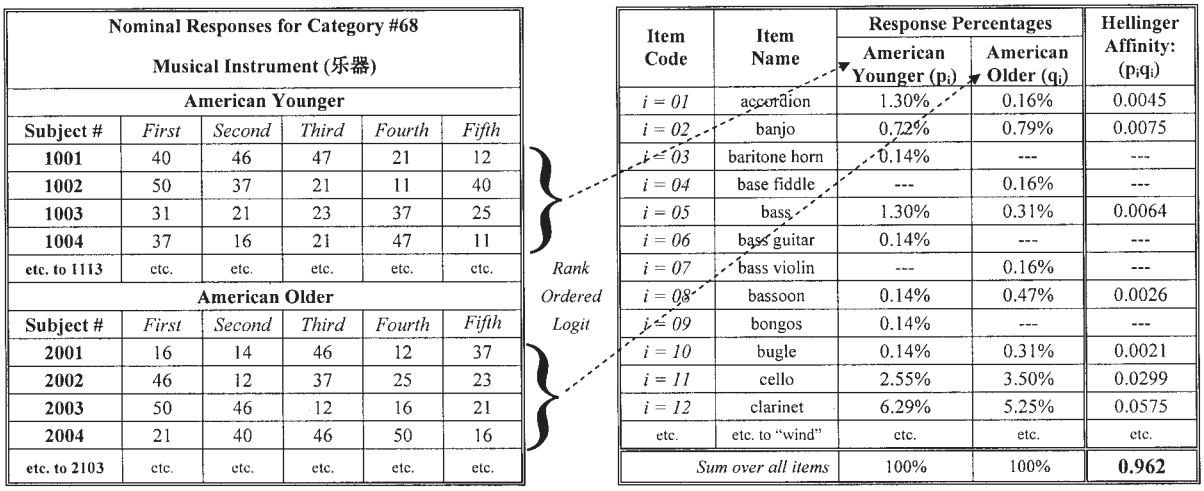
Categories varied widely in terms of the number of unique items generated across the four groups. For example, there were only 40 total item responses across four groups for "farm animal" and 44 for "part of a face," but there were 270 for "burial place," 277 for "toy," and 419 for "reference book." Although the data show that there is indeed a relationship between the number of unique items in a category and measures of between-group similarity, this comes about for a simple reason: Categories have many items, largely *because* the sets named by the various groups displayed minimal overlap. That is, each culture-by-age group views the category differently from the others, and so many items are generated. Unlike those of prior studies, the methodology we present next systematically accounts for this fact.

Analyses

Freely Generated Response Data and Measuring Nominal Distance

We wished to rigorously measure the extent to which any two groups agree on item frequencies within a category. For example, we might ask whether younger Chinese and younger Americans name different types of trees with similar (distributional) item frequencies. Doing so required a method to determine how well two nominal[6] distributions accord, presenting unique inference problems (Agresti, 1990, 1992). To see why this might be so, consider three qualities intrinsic to these data.

[6] To avoid confusion with categories in the study, we refer to categorical data and methods as *nominal.*



Nominal Responses for Category #68

Musical Instrument (乐器)

American Younger

Subject #	*First*	*Second*	*Third*	*Fourth*	*Fifth*
1001	40	46	47	21	12
1002	50	37	21	11	40
1003	31	21	23	37	25
1004	37	16	21	47	11
etc. to 1113	etc.	etc.	etc.	etc.	etc.

American Older

Subject #	*First*	*Second*	*Third*	*Fourth*	*Fifth*
2001	16	14	46	12	37
2002	46	12	37	25	23
2003	50	46	12	16	21
2004	21	40	46	50	16
etc. to 2103	etc.	etc.	etc.	etc.	etc.

Rank Ordered Logit

Item Code	Item Name	Response Percentages: American Younger (p_i)	Response Percentages: American Older (q_i)	Hellinger Affinity: (p_iq_i)
i = 01	accordion	1.30%	0.16%	0.0045
i = 02	banjo	0.72%	0.79%	0.0075
i = 03	baritone horn	0.14%	---	---
i = 04	base fiddle	---	0.16%	---
i = 05	bass	1.30%	0.31%	0.0064
i = 06	bass guitar	0.14%	---	---
i = 07	bass violin	---	0.16%	---
i = 08	bassoon	0.14%	0.47%	0.0026
i = 09	bongos	0.14%	---	---
i = 10	bugle	0.14%	0.31%	0.0021
i = 11	cello	2.55%	3.50%	0.0299
i = 12	clarinet	6.29%	5.25%	0.0575
etc.	etc. to "wind"	etc.	etc.	etc.
	Sum over all items	100%	100%	**0.962**

Figure 1. Data analysis process: Rank-ordered logit model and Hellinger Affinity (.962) for Category 68, musical instrument (乐器). Depicted are the first through fifth item responses for 4 younger American and 4 older American participants. For example, Subject 1001's item codes {40, 46, 47, 21, 12} correspond to written, verbal responses {saxophone, trumpet, tuba, flute, clarinet}, respectively. The rank-ordered logit model transforms such ranked data for each group into underlying response probabilities, as indicated, for each of the musical instruments in the category. Finally, the Hellinger Affinity (HA) can be calculated from these response probabilities and measures the degree of item response similarity. For this comparison, the HA of .962 indicates exceptionally strong agreement between the older American and younger American groups in their musical instrument item responses.

Partial ranking. In a given category, observations consist of each participant's top five items. The number of possible partial orderings of this type is enormous. Further, the ranked structure of the data must be accounted for, so that each item counts more than those listed beneath it and all listed items count more than those not listed.

Sparseness and nonmentioned items. Many items are mentioned only once, so for many (perhaps most) cells, observed frequencies are small, even zero. Even more problematic, some items mentioned by one group are never mentioned at all by the other(s).

Items versus observations. The number of items in a category is not known in advance but will grow with the sample size. The set of all items that *can* be mentioned may be of unreasonable size, making it difficult to construct the space of possible responses, probability distributions over it, and "distance" between empirical frequency counts.

A great deal of research has addressed whether a number of samples appear to be drawn from the same nominal distribution. Generally speaking, measures (such as the Pearson chi-square and the likelihood ratio statistic G^2) that rely on the approximate normality of the underlying Poisson cell count data require expected cell frequencies to be roughly five or more (Cressie & Read, 1984; Koehler & Larntz, 1980). Such conditions seldom hold for freely generated data. Even if one were to ignore the rank-ordered nature of the individual-level data (e.g., using only the first response for each participant), so-called sparse data entail problems with overdispersion, leading to grossly overstated significance levels for standard chi-squared-based tests of between-group differences (Haberman, 1988). And even standard methods for sparse data can falter when the number of items grows without bound (Zelterman, 1987).

We were therefore faced with two separate problems: that of measuring a degree of association between two observed nominal distributions (an *affinity measure*) and that of accounting for the rank-ordered nature of the data (a *ranked-data model*). Although the Pearson and Spearman correlations are familiar affinity measures, the assumptions they require are typically not met for nominal data, for sparse data, or for rank-ordered data; here, we have all three. So, it is impossible to know whether inferences based on those correlations would be valid for our data and, moreover, which of these three aspects of the data might cause them to go awry. To see that this is so, consider the three following thought examples, each highlighting a potential problem in applying correlation-type measures to rank-ordered nominal data: the need for ad hoc rules of application; their penalization of order more than lack of overlap; and their equal weight on all items, regardless of frequency. We look at each, briefly, in turn.

Rules of application. First, suppose that younger Chinese and older Chinese name 10 items very frequently between them but their frequencies are in reverse order, so that the most common response by the former was 10th among the latter, and so forth. If we restrict our analysis to these 10 items, the Pearson correlation (or even the Spearman rank-order correlation) would indicate a nearly perfect negative relationship, leading to a strong conclusion of between-group differences for that category. Suppose, however, that another hundred items mentioned far less frequently now occupied ranks 11 through 110 on both lists, in identical order. A correlational measure would now be quite near 1, suggesting an opposite conclusion. *Which* items (i.e., how many) one includes in the analysis therefore has a decisive, and unwarranted, effect on the inferred strength of item distribution similarity. To help overcome such problems, one might enact ad hoc rules to determine the number of items to use for calculation, a procedure made more

difficult, and less defensible, when item lists differ across subject groups. To our knowledge, the statistical literature is silent on how to choose such a number, even when list sizes are known in advance.

Order versus overlap. Consider again two groups, for each of which we have retained the top 10 items, which happen to be the same. If their frequencies are identical, the Pearson correlation is 1; if the frequencies are identical but in reverse order, the correlation is −1. However, if the two groups had mentioned *no* items in common, the resulting correlation would be close to zero,[7] or at the very least, somewhere near the center of the −1 to 1 scale. Intuitively, a complete lack of item list overlap between groups should indicate a radically different view of the category and thus the lowest possible affinity value. Correlational measures are most strongly influenced in the negative direction when the *order* of items on which the two groups *agree* happen to be reversed, not when the groups disagree completely on what those items are in the first place. The method we apply does not share this property.

Weighting based on frequency. Last, standard correlational measures allow equal influence (leverage) to higher frequency and lower frequency items, a counterintuitive feature when measuring whether *most* members of two groups agree. When an item goes nearly unmentioned by each of two groups, it increases the Pearson correlation; yet there are potentially limitless numbers of such items that can enter into the computation, unless some arbitrary rule is used to limit them a priori. An appropriate affinity measure should place emphasis on items in direct proportion to their frequency of mention.

These features, in our view, call standard correlation-based measures very much into question as affinity measures for freely generated, sparse data of the type analyzed here.[8] Further, it is unclear how one would include the partially ranked nature of the individual-level data, to which we turn next.

Ranked Nominal Data

When analyzing rank-ordered data, it is possible to disregard all but the first item mentioned or to count only whether an item was mentioned at all, simply ignoring the ranking information. We believe, however, that it is important to retain the ranked nature of participant responses, for several reasons. First, participants were requested to list five responses for each category, in the order they came to mind. Data models that ignore this crucial feature presume that later mentioned items—which some participants may have strained to come up with—represent the category as well as earlier ones. Second, in many categories there were pairs of items that appeared on participants' lists with nearly equal frequency, though one typically preceded the other, indicating greater ease of recall or degree of prototypicality. Ignoring this consistent feature of the ranked data, we believe, would have obscured the simple fact that many categories had strong exemplars that appeared near the top of all participants' lists. Third, analyses ignoring the rank-order information can produce biased results, simply because relevant structural information is deliberately ignored. For example, among the older Chinese, in the category "type of metal," iron appeared on the lists of 86% of participants and appeared first for 18%; silver, by contrast, appeared on 74% of the lists and was never listed first. One would feel confident in concluding that iron should rank well ahead of silver. However, a closer look at the data makes the opposite case: On lists where *both* appeared, silver was nearly always ranked higher. The reason for this is that gold was by far the most likely first choice, and in the 57% of cases where it appeared in first position, silver was typically second (in fact, the modal rank of silver was 2). This case was hardly anomalous; the data speak plainly that first occurrence and overall occurrence rates[9] fail to paint an accurate portrait of the distribution of item responses in many categories. The rank-ordered logit model, described in the next section, presents a parsimonious and intuitively appealing way of accounting for these item distributions.

The Rank-Ordered Logit Model

In line with much prior work in economics, we adopt the rank-ordered (or "exploded") logit model to analyze participants' item response data. This model has an attractive interpretation in terms of underlying utility (McFadden, 1974), though we present it in more heuristic terms. An accessible treatment of partially ranked data methods can be found in Critchlow's (1985) monograph, and empirical illustrations of the rank-ordered logit model appear in Skrondal and Rabe-Hesketh's (2003) examination of multilevel data, in Koop and Poirier's (1994) analysis of voting preferences, and in the classic expository article of Chapman and Staelin (1982).

In the present application, each participant's set of choices in a given category is viewed as a draw of size five from a nominal distribution (i.e., a multinomial process), with unknown, group-specific frequencies. For each category, the model reconstructs those frequencies, for each culture-by-age group, from participants' ordered choice data, as illustrated in Figure 1. Because we seek a *univariate* measure (i.e., a single number) of how well two categories accord, it is necessary to "derank" the ranked data while retaining the information embodied in the rankings to the greatest extent possible. The rank-ordered logit model is uniquely suited to this task.

An example helps clarify how the model deranks participants' data while remaining faithful to the item response distribution. Assume, for purposes of illustration, that the category in question is one not used in the study, "suits of playing cards," and that only four responses were generated by participants: {hearts, spades, diamonds, clubs}, with respective probabilities {.4, .3, .2, .1}. In practice, these probabilities are not known in advance, and the purpose of applying the model is to reconstruct them from the ranked data. If participants are asked to choose two from among

[7] This is true only if we exclude items not appearing on both lists. If one includes any items appearing on either list, the Pearson correlation would be (mildly) negative, typically in the −.30 to −.50 range, depending on particulars of the item distribution.

[8] We do not mean to suggest that the use of such measures in prior research led to incorrect inferences; this would require, at the very least, a full reanalysis of the individual-level data in question. Extensive simulations (available from the authors) indicate, however, that correlation-based measures are highly sensitive both to decisions researchers must make in how to apply them and to distributional features of item responses in a category. The methods developed here do not share these sensitivities.

[9] Both first occurrence and overall occurrence rates are calculated for all 105 categories and each of the four culture-by-age groups on the project Web site, along with other summary statistics.

the four suits, where order matters, the probability for any such subset is then readily computed. Given the unconditional probabilities, {.4, .3, .2, .1}, the most likely subset consists of {hearts, spades}, in that order: The probability for hearts is .4, and the (conditional) probability of choosing spades from among those remaining is .3/[.3 + .2 + .1]. Thus, the joint probability of the ordered subset {hearts, spades} is (.4)(.3)/(.3 + .2 + .1) = 12/60. An analogous calculation shows that order matters: The probability of {spades, hearts} is (.3)(.4)/(.4 + .2 + .1) = 12/70, a smaller number. Calculations for any overall set size, any number of selected items, and any order are exactly analogous. From these probabilities, a sample likelihood is constructed; the log-likelihood contribution for a single participant whose top five choices have probabilities $\{p_1, \ldots, p_5\}$ is

$$LL = \sum_{k=1}^{5}[\log(p_k) - \log(1 - \sum_{j=1}^{k-1} p_j)].$$

Standard Newton–Raphson gradient search methods readily yield maximum likelihood estimates for the unknown, underlying item probabilities for each culture-by-age group in each category.[10] It is these deranked response probabilities, as shown in Figure 1, that are used in subsequent analysis.

The proposed methodology, and all prior ones, can be criticized for not modeling data across categories and for assumptions of within–culture-by-age-group homogeneity. With regard to terminology, note that data across categories are purely nominal, whereas data within categories (for a particular respondent) are of the "partially ranked" type, in the sense of Critchlow's (1985) classic monograph. Thus, the data record for each respondent spans multiple categories and includes items from different, category-specific sets. Modeling such data across categories, considering that no scales were used, is infeasible even through state-of-the-art Bayesian methods. Extant models that account for different forms of response heterogeneity (e.g., Rossi, Gilula, & Allenby, 2001) require that participants respond across categories using some type of explicit scale for each one. It is further well beyond current methods to offer an account of within-group heterogeneity. A typical approach, for example, appeals to some form of mixture model—allowing there to be different segments or classes within each culture-by-age group—but this, too, is ruled out by the sparse, nominal nature of the categories.

Measuring Differences With Sparse Nominal Distributions: The Hellinger Affinity

Among the main goals of the present study was to create a taxonomy of categories for use in cross-cultural research. That is, we wished to determine whether, say, older Chinese and older Americans differ in their item responses for "article of clothing" more than for "source of energy" and to compare all categories for all groups in a similar manner. To do so required a measure of which categories most differ in their item response profiles across the four culture-by-age groups. Prior research has mainly made use of measures reliant on a regression framework. Although these have the benefit of familiarity, they rest on assumptions that demonstrably do not hold for nominal data (Liebetrau, 1983). Instead, we relied on measures common in natural language processing and neural net studies, so-called divergence and distance metrics, which are naturally suited to nominal distributions. In particular, we used the Hellinger Affinity (HA) and the Jensen–Shannon entropy (JS), which each have the appealing property of lying on a 0–1 scale, with higher values indicating a greater degree of nominal overlap, thus resembling R^2, U^2, or the likelihood-ratio statistic. Fine expository accounts of the theory and use of distance measures (such as the HA) and divergence measures (such as the JS), as well as their interconnections, can be found in Lee (1999) and Pollard (2001).

The HA has a natural interpretation in terms of the distance between two frequency distributions (specifically, the cosine of the angle between their square roots). Given n items, if the observed proportions in each of two groups are given by $(p_1, p_2, \ldots, p_n)$ and $(q_1, q_2, \ldots, q_n)$, then

$$HA(p,q) = \sum_{i=1}^{n} \sqrt{p_i q_i}.$$

It takes on its largest value, HA = 1, when the distributions are identical, and its lowest value, HA = 0, when there is no overlap whatsoever. The JS measure relies on the concept of statistical divergence, which, though less intuitive, is nonetheless foundational.[11] We included the JS measure in all project files both for completeness and as a check on reliability, though we relied primarily on the HA for subsequent comparisons; substantive results were nearly identical.

It is important to realize that both the HA and JS measures are indeed that: measures, not statistical tests.[12] Each is rather like an R^2—in that it summarizes the intrinsic strength of the nominal relationship—and not like the associated F test, which assesses whether the "real" population relationship is exactly zero and

[10] All numerical analysis was done in MATLAB; the programs written for this purpose are available on request.

[11] The JS measure depends on the well-known Kullback–Leibler divergence, defined for any two sample probability distributions as

$$D[p|r] = \sum_{i=1}^{n} p_i \log(p_i/r_i).$$

If r is the average of p and q, then $JS(p, q) = (D[p \mid r] + D[q \mid r])/2$. It is thus the mean divergence of two distributions from their own average. See Lee (1999) for additional detail.

[12] Because none of the divergence or distance measures has a standard distribution under the null hypothesis of equal between-group item frequencies, they can be used for testing only through computationally intensive resampling methods. Although similar tests are possible using the likelihood-ratio statistic or other standard nominal analysis tools, we stress that these are known to grossly overstate significance levels for sparse data. There are no general methods for such situations (Agresti, 1990, 1992), and tests based on correlational measures are known to be highly misleading (Koehler & Larntz, 1980). In fact, extensive analyses using both resampling and the Friedman-Rafsky (1983) test show that essentially all cross-age and cross-culture category differences are "significant"; this merely reflects that there is a great deal of data, hundreds of participants each offering five pieces of information, so even small, unimportant affinity differences give rise to p values under .05. These results are available on request.

which is strongly influenced by sample sizes. It is important to note that the HA and JS measures therefore offer a sense of how substantial group differences really are, irrespective of the quantity of data, and can be compared across groups of participants, categories, and even studies.

In short, among other attractive properties, affinity scores (a) are rigorously defensible measures of nominal agreement; (b) are readily interpretable, on a unit scale, from no overlap to perfect overlap; (c) weight frequent item responses more heavily than infrequent ones; (d) allow comparison of *degree* of difference between any two of the culture-by-age groups; and, perhaps most important, (e) require no arbitrary cutoffs or decisions in their application and are in fact very simple to calculate. Although other methods possess some of these properties—for example, the free-list salience index used in cultural anthropology (Smith, 1993)—the one we developed is the only method we know of that posits underlying theoretical measurement models: random utility theory (for the rank-ordered logit) and geodesic distance from differential geometry (for the HA measure).

For each of the 105 categories, five measures were calculated for every response item generated by each of the four culture-by-age groups: (a) response probability (deriving from the logit model); (b) frequency of overall occurrence (percentage); (c) frequency of first occurrence (percentage); (d) mean rank order; and (e) weighted rank order.[13] In addition, the number of unique response items generated by each culture-by-age group was calculated for each of the 105 categories.

Comparison of Affinity Scores Across Culture and Age

A primary purpose of this study was to introduce the archive of culture-by-age group category norms and explore how researchers might best put it to use. Because it contains data on 105 separate categories, each involving hundreds of item responses across four culture-by-age groups and in two languages, the number of potential comparisons is vast, and we cannot completely characterize them here. Rather, we highlight the most important comparisons and resulting conclusions, relegating all remaining comparisons to the project Web site.

Affinity scores suggest that 13 categories (12% of 105 categories) have roughly equivalent category responses across all four culture-by-age groups and are thus suitable for use in cross-cultural studies: time unit, internal organ, tree (part), season, color, face (part), farm animal, mathematical operation, metal, fruit, chemical element, human body (part), and wild animal. All placed among the top 15 most equivalent categories (out of 105 categories) for each cross-cultural comparison (i.e., younger and older) and scored very high in cross-age comparisons (with affinities above .90). Several other categories—writing implement, time of day, bicycle (part), meat, insect, and reading material—fared almost as well but showed more substantial item deviations and would need to be used more judiciously. For example, if researchers were able to restrict possible item responses, several other categories could be made to have similar frequency distributions among the items explicitly allowed. Affinity scores are readily computed on these conditional distributions, from frequency data available on the project site.[14]

The main results are well summarized by Table 2, which lists the top 25 categories, ordered by affinity, for each culture-by-age group. A number of generalities are readily apparent. First, there is a far greater degree of confluence across ages than across cultures: Participants within a culture, both younger and older, viewed the same categories in a similar manner, whereas those in the same age group (across cultures) generally differed more. Because of this, a far greater number of categories are suitable for cross-age (within a culture) than for cross-cultural studies.

Second, some of the categories most agreed on in one culture performed less well in the other. Perhaps stereotypically, the single best affinity score among Americans (younger vs. older; column 1) was that for dairy product, its .985 value indicating nearly perfect overlap. By contrast, dairy product finished 35th (out of 105) among Chinese (younger vs. older; column 2; HA = .897). The same is true, for example, of royalty (member of), type of waterway, and others. This raises an important, and subtle, point: Although there is surely an overall main effect of intrinsic category similarity, generally speaking, it is not nearly strong enough to explain within-culture agreement. Even a cursory look at the category orderings in Table 2 makes clear that effects of culture are simply far more potent than those of age in determining how participants agree on item categorization. Interestingly, some categories emerged as uniformly inappropriate for any studies at all: Tropical fish, herbal medicine, and folk art performed quite poorly across age and exceptionally so across culture.

It does appear, then, that some categories are suitable in any culture-by-age setting, whereas others are decidedly unsuitable. One might naturally question whether there are some qualities the suitable categories possess that others do not. To address this, we parceled the categories into 14 larger types and considered the overall pattern of affinity scores.[15] Two types contained only cross-culturally suitable categories: *measures and abstract properties* (color, compass direction, mathematical operation, shape,

[13] Functions of rank, such as the mean and weighted rank, were computed as a form of shorthand summary and to accord with prior studies. Because ranks are not interval scaled, the mean cannot be taken to reflect central tendency, and even a single outlier can greatly affect it. The weighted rank position measure for each response, a form of the well-known Borda–Kendall count (Cook & Seiford, 1982; Felsenthal, Moaz, & Rapoport, 1993), was calculated as follows: [(5 × frequency of first occurrence) + (4 × frequency of second occurrence) + . . . + (1 × frequency of fifth occurrence)]/15. By definition, these sum to 100% across all items. By contrast, mean rank averages an item's position, given that it appears on a participant's list. Thus, an item appearing on few lists but in a high position will have a very high mean rank, so the mean rank figures reported here and in prior studies must be interpreted with caution. Generally speaking, weighted rank is a more reliable guide to item distribution.

[14] Item-level data for all 105 categories are available from the project site, organized by culture and age. HA values and JS divergences for pairwise comparisons of the four culture-by-age groups and all 105 categories are presented in the Appendix. Further detail and comparisons appear on the project Web site.

[15] This information is available at the project Web site under "Categories by Category Type" (see Footnote 4 for URL). The 14 category types were activities and entertainment, animals, buildings and locations, food and beverages, household items, measures and abstract properties, medical and human-related, natural phenomena, personal effects and items, plants and trees, science and technology, social roles and constructs, tools and materials, and transportation modes and vehicles. All categorical information and measures on the site are keyed to this typology as well.

Table 2
Category Equivalence: All Culture-by-Age Groups, Top 25 Categories

American: younger vs. older		Chinese: younger vs. older		Younger: American vs. Chinese		Older: American vs. Chinese	
Category name	Hellinger Affinity	Category name	Hellinger Affinity	Category name	Hellinger Affinity	Category name	Hellinger Affinity
Dairy product	.985	Compass direction	.990	*Time* (*unit*)	.962	*Organ* (*internal*)	.942
Time (unit)	.983	Face (parts of)	.988	*Organ* (*internal*)	.934	*Mathematical operation*	.941
Royalty (member)	.982	Animal (farm)	.985	*Tree* (*parts of*)	.928	*Color*	.927
Season	.976	Meat	.983	*Season*	.907	*Animal* (*farm*)	.905
Organ (internal)	.974	Color	.976	*Color*	.904	*Tree* (*parts of*)	.887
Face (parts of)	.972	Fruit	.972	*Face* (*parts of*)	.901	*Face* (*parts of*)	.880
Room in house	.965	Animal (four-footed)	.964	*Animal* (*farm*)	.895	*Season*	.880
Musical instrument	.962	Tree (parts of)	.961	*Mathematical operation*	.888	*Fruit*	.866
Fruit	.962	City	.961	*Metal*	.853	*Metal*	.828
Mathematical operation	.957	Footwear	.957	Writing implement	.841	*Time* (*unit*)	.817
Animal (farm)	.956	Weather	.950	*Fruit*	.839	*Animal* (*wild*)	.809
Color	.954	Organ (internal)	.949	*Chemical element*	.830	Meat	.780
Animal (four-footed)	.951	Vegetable	.946	*Human body* (*part*)	.794	Compass direction	.776
Storm	.947	Mathematical operation	.944	*Animal* (*wild*)	.776	*Chemical element*	.774
Compass direction	.944	Writing implement	.936	Precious stone	.773	*Human body* (*part*)	.761
Appliance (major)	.940	Precious stone	.934	Bicycle (parts of)	.761	Furniture	.744
Building (religious)	.939	Medical specialty	.934	Time of day	.754	Time of day	.739
Waterway	.938	Bird	.934	Building (religious)	.753	Animal (four-footed)	.726
Animal (wild)	.937	Season	.933	Science	.748	Writing implement	.719
Time of day	.936	Shape	.931	Insect	.747	Singing voice	.684
Weather	.936	Insect	.930	Shape	.736	Insect	.680
Singing voice	.935	Appliance (major)	.926	Musical instrument	.731	Energy (source)	.671
Vegetable	.932	Nut	.925	Meat	.702	Bicycle (parts of)	.665
Military title	.932	Jewelry	.920	Reading material	.690	Energy	.664
Bicycle (parts of)	.930	Time of day	.917	Nut	.686	Reading material	.661
Spearman rank-order correlation = .584				Spearman rank-order correlation = .927			

Note. Italics denote categories best suited for cross-cultural studies. All such categories are appropriate for cross-age studies. Correlations are computed across all 105 categories. For both correlations (.584 and .927), and for the difference between them, Fisher's *r*-to-*z* transformation yields $p < .0001$.

time unit, and time of day) and *natural phenomena* (natural earth formation, precious stone, season, storm, waterway, and weather); other suitable categories fell predominately into *medical and human related* (e.g., face [part], human body [part], and internal organ) and *plants and trees* (e.g., fruit and tree [part]). By contrast, several types contain mainly categories ill suited cross-culturally: *activities and entertainment* (e.g., reference book, circus act, and folk art), *buildings and locations* (e.g., burial place, city, and human dwelling), *household items* (e.g., small kitchen appliance, bathroom fixture, and kitchen utensil), and *personal effects and items* (e.g., footwear, jewelry, and toy).

Although our study was not designed specifically to address such issues, some generalizations based on this typology are appropriate. The suitable categories are those that deal with canonical properties, those deriving from nature or extremely broad social convention (as in the case of units of time and measure, overseen by cross-national standards groups). That there is general agreement on which items best represent these categories is, in some sense, unsurprising. The converse concerns the unsuitable categories, which deal with local reference, culture, and custom. The categories with the lowest affinity scores—specifically, herb, folk art, city, and mythical being—are culturally or locally dependent in a transparent manner. The natural–cultural (or global–local) dichotomy holds well for those types containing both suitable and unsuitable categories: For *animals*, farm animal and wild animal had very high affinity scores, whereas breed of dog, fish, and tropical fish had among the lowest; for *food and beverages*, the natural categories (e.g., nut, type of meat) had high affinity scores, whereas the local ones (e.g., bread, candy, dairy product) had among the lowest.

The distinction between natural (or global) and cultural (or local) is not hard and fast, but nonetheless, it has broad applicability. For example, a day or the concept of subtraction does not differ appreciably from culture to culture or place to place, unlike, say, mythological figures. There are, however, categories for which the natural–cultural distinction is somewhat blurry. For instance, owing to the homogenization of large-scale food production, "farm animal" hardly varies across culture, although it may have a century ago. By contrast, "tropical fish" changes dramatically from one location to another and might be expected to have a far stronger local component. Overall, although our examination is informal, affinity scores nevertheless agree with intuition: that what philosophers call "natural kinds" offer a better basis for culturally neutral stimuli.[16]

Affinity scores can be used to address an intriguing metaquestion. Among the younger participants, one can order all categories

[16] This broadly accords with the studies of cross-linguistic perceptual terms, such as those for color (Kay et al., 1997). Lin, Luo, MacDonald, and Tarrant (2001) studied color-naming terms in Mandarin- and in English-speaking participants, finding "close agreement." Our own data are similar, with color having among the highest affinity scores: over .90 across cultural for both age groups and over .95 across age for both culture groups.

by how similar the American and Chinese participants found them, and one can do the same for the older American and older Chinese participants (as in Table 2). How similar are these orderings? That is, do the younger people order the categories (by item distribution) the same way the older people do, across cultures? The third and fourth columns of Table 2 show clearly that the categories for which the younger Chinese and younger Americans have high levels of agreement are nearly identical to those strongly agreed on by their older counterparts. The Spearman rank-order correlation for this comparison is .927, indicating a startling degree of accord in which categories are viewed similarly. It is important to realize that the *ordering* of categories by affinity score ignores their absolute levels and speaks in a nonparametric way as to how well the orderings accord. By contrast, one can perform the analogous procedure within a culture group, that is, order categories by how well younger Chinese agree with older Chinese, do the same for younger Americans versus older Americans, and compute the Spearman correlation. Although the resulting value of .584 is still fairly high, it suggests that there is far greater agreement about categories within culture than within age.[17] Given the Spearman values, it seems fair to conclude that age adds relatively little, on the margin, to our understanding of cross-culture category agreement, whereas culture does add to even the considerable understanding based on age alone.

Discussion

One of the primary purposes of this study was to obtain norms on a set of categories that are suitable for use in studies of cross-cultural or cross-age differences in cognitive performance. Researchers will now have access, before any data are collected, to a list of categories that generalize across cultures and age, as well as a list of those that are different across cultures and age. The second purpose was to provide an appropriate statistical framework for analysis of the data, taking into account both the relative frequency and order of production of item responses generated within each category.

The study revealed a relatively large number of categories with similar sets of response items when younger versus older adults were compared within each culture. For the Chinese in our study, 85% of the categories (89 out of 105) had strong agreement in terms of the items represented across the age groups (i.e., affinity scores at least .90); for Americans, 91% of the categories (96 out of 105) had strong agreement across the two age groups. This generally accords with findings by Howard (1980), who reported similar response frequencies for the two age groups across all 21 categories tested. However, Howard carefully paired this finding with an often overlooked caveat: that the categories included in the study were selected on the basis of being the least likely to yield age differences. Howard, in fact, explicitly cautioned against generalizing findings of high correspondence to other categories. Although review articles and books on cognitive aging frequently suggest that there is age invariance in semantic organization and processes (e.g., Light, 1991), direct empirical evidence supporting this view is scant.

Our findings further challenge the notion of complete age invariance in terms of category representations. Although we found high agreement within each culture for a majority of the categories tested, it clearly cannot be assumed that all category norms are stable across the lifespan. There are substantial age differences for categories in domains for which familiarity or experience levels may differ between younger and older groups (e.g., disease, herbal medicine). This is consistent with recent developments in cognitive science. Specifically, there is emerging evidence that the brain exhibits plasticity well into advanced age and that experiences lead to changes in neural representation across the lifespan (e.g., Squire & Kandel, 1999). To the extent that life experiences differ across age groups, one would expect little overlap in the ways category information related to those experiences are mentally represented in younger and older adults within a culture.

The above reasoning applies all the more to comparisons across cultures. To the degree that systematically different life experiences yield systematically different cognitive contents and processes, one would expect considerable variability between American and Chinese individuals, in both the number and types of normative responses generated. Imposing a rather strict rule that all culture-by-age-group comparisons reflect high affinity scores (.90 or greater), relatively few categories—only 13 of the 105 tested—appear to have stable agreement in terms of the generated instances across all four culture-by-age groups. For researchers investigating cross-cultural differences in cognitive performance limited to younger adults, 6 additional categories (for a total of 19) are suitable. Limiting one's purview to older adults across the two cultures allows for 3 additional categories (for a total of 16).

Our results thus indicate that there are substantial differences in category norms between the cultural groups and smaller differences across age groups within each culture. Hence, we suggest that unequivocal interpretations or conclusions with respect to cross-cultural differences, and to a lesser extent age-related differences, in cognition can be made only through recourse to appropriate, standardized stimuli. The corpus of data now available should therefore be used as a guide by researchers for choosing appropriate categories, depending on the purpose of the planned studies. For example, in designing studies to examine cross-cultural differences in category fluency or knowledge, researchers will know a priori which categories yield equivalent responses, and which completely different responses, for Chinese and Americans. These data should also serve to promote more research on a topic of burgeoning interest to cognitive aging researchers: how cultural and life experiences shape cognitive contents and functioning across the lifespan.

[17] Using Fisher's *r*-to-*z* transformation, we were able to test whether the sample values of .927 and .584 indicated an ordinal relationship ($z = 6.76$ and 16.54, respectively), as well as their difference ($z = 6.92$). All are highly significant, and the last suggests that category orderings based on culture agree significantly more than those based on age.

References

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science, 7,* 131–177.

Avgerinos, A., & Theofilos, T. (1995). *A linguistic approach to information retrieval.* Doctoral dissertation, University of Nijmegen, Nijmegen, the Netherlands. Retrieved June 15, 2004, from http://citeseer.nj.nec.com/arampatzis96linguistic.html

Balota, D. A., Black, S. R., & Cheney, M. (1992). Automatic and atten-

tional priming in young and older adults: Reevaluation of the two-process model. *Journal of Experimental Psychology: Human Perception and Performance, 18,* 485–502.

Battig, W. F., & Montague, W. E. (1969). Category norms of verbal items in 56 categories: A replication and extension of the Connecticut category norms. *Journal of Experimental Psychology, 80,* 1–46.

Birren, J. E., & Fisher, L. M. (1995). Aging and speed of behavior: Possible consequences for psychological functioning. *Annual Review of Psychology, 46,* 329–353.

Boccardi, M., & Cappa, S. F. (1997). Valori normative di produzione categoriale per la lingua italiana [Normative values of categorical production for the Italian language]. *Giornale Italiano di Psicologia, 24,* 425–436.

Bond, M. H. (1991). *Beyond the Chinese face: Insights from psychology.* New York: Oxford University Press.

Brown, W. P., & Davies, G. M. (1976). Studies in word listing: Testing for group differences in category norms. *Irish Journal of Psychology, 3,* 87–120.

Casey, P. J., & Heath, R. A. (1988). Category norms for Australians. *Australian Journal of Psychology, 40,* 323–339.

Chan, J. (1996). Chinese intelligence. In M. H. Bond (Ed.), *The handbook of Chinese psychology* (pp. 93–108). New York: Oxford University Press.

Chapman, R. G., & Staelin, R. (1982). Exploiting rank ordered choice set data within the stochastic utility model. *Journal of Marketing Research, 19,* 288–301.

Cohen, B. H., Bousfield, W. A., & Whitmarsh, G. A. (1957). *Cultural norms for verbal items in 43 categories.* Studies on the Mediation of Verbal Behavior (Tech. Rep. No. 22). Storrs, CT: University of Connecticut.

Cook, W. D., & Seiford, L. M. (1982). On the Borda–Kendall consensus method for priority ranking problems. *Management Science, 28,* 621–637.

Cressie, N., & Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society (Series B, Methodological), 46,* 440–464.

Critchlow, D. E. (1985). *Metric methods for analyzing partially ranked data* (Lecture Notes in Statistics, Series 34). New York: Springer-Verlag.

Felsenthal, D. S., Moaz, Z., & Rapoport, A. (1993). An empirical evaluation of six voting procedures: Do they really make any difference? *British Journal of Political Science, 23,* 1–27.

Fiske, A., Kitayama, S., Markus, H. R., & Nisbett, R. E. (1998). The cultural matrix of social psychology. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The handbook of social psychology* (pp. 915–981). New York: McGraw-Hill.

Friedman, J. H., & Rafsky L. C. (1983). Graph-theoretic measures of multivariate association and prediction. *Annals of Statistics, 11,* 377–391.

Goikoetxea, E. (2000). Frecuencia de produccion de las respuestas a categorias verbales en ninos de primaria [Production frequency for verbal items in 52 verbal categories for school children]. *Psicologica, 21,* 61–89.

Haberman, S. J. (1988). A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. *Journal of the American Statistical Association, 83,* 555–560.

Hampton, J. A., & Gardiner, M. M. (1983). Measures of internal category structure: A correlational analysis of normative data. *British Journal of Psychology, 74,* 491–516.

Hasselhorn, M., Jaspers, A., & Hernando, M. D. (1990). Typizitaetsnormen zu zehn Kategorien fuer Kinder von der Vorschule bis zur vierten Grundschulklasse [Category norms for verbal items in ten categories for children in kindergarten and in Grades 1–4]. *Sprache & Kognition, 9,* 92–108.

Henik, A., & Kaplan, L. (1988–1989). Category content: Findings for categories in Hebrew and a comparison to findings in the US. *Israel Journal of Psychology, 1,* 104–112.

Howard, D. V. (1980). Category norms: A comparison of Battig and Montague (1969) norms with the response of adults between the ages of 20 and 80. *Journal of Gerontology, 35,* 225–231.

Huang, H. S., & Hanley, J. R. (1994). Phonological awareness and visual skills in learning to read Chinese and English. *Cognition, 54,* 73–98.

Huang, S., & Jia, W. (2000). The cultural connotations and communicative functions of Chinese kinship terms. *American Communication Journal, 3.* Retrieved June 15, 2004, from http://acjournal.org/holdings/vol3/Iss3/curtain.html

Hupbach, A., & Mecklenbräuker, S. (1998). Typizitätsnormen zu neun Kategorien für Kindergartenkinder zweier Altersstufen [Category norms for verbal items in nine categories for younger and older preschoolers]. *Sprache & Kognition, 17,* 41–50.

Jeng, C., Lai, M., & Liu, I. (1973). Category norms in Chinese and English from bilingual subjects. *Acta Psychologica Taiwanica, 15,* 81–153.

Kausler, D. H. (1991). *Experimental psychology, cognition, and human aging* (2nd ed.). New York: Springer-Verlag.

Kay, P., Berlin, B., Maffi, L., & Merrifield, W. (1997). Color naming across languages. In C. L. Hardin & L. Maffi (Eds.), *Color categories in thought and language* (pp. 21–56). Cambridge, UK: Cambridge University Press.

Koehler, K. J., & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association, 75,* 336–344.

Koop, G., & Poirier, D. (1994). Rank-ordered logit models: An empirical analysis of Ontario voter preferences. *Journal of Applied Econometrics (October–December),* 369–388.

Lee, L. (1999, June 22–27). *Measures of distributional similarity.* Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. Retrieved June 15, 2004, from http://www.cs.cornell.edu/home/llee/papers.html

Liebetrau, A. M. (1983). *Measures of association* (Quantitative Applications in the Social Sciences, Series No. 32). Newbury Park, CA: Sage.

Light, L. L. (1991). Memory and aging: Four hypotheses in search of data. *Annual Review of Psychology, 42,* 333–376.

Light, L. L., & Singh, A. (1987). Implicit and explicit memory in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 531–541.

Lin, H., Luo, M. R., MacDonald, L. W., & Tarrant, A. W. S. (2001). A cross-cultural colour-naming study: Part I. Using an unconstrained method. *Color Research and Application, 26,* 40–60.

Lin, P. J., & Schwanenflugel, P. J. (1995). Cultural familiarity and language factors in the structure of category knowledge. *Journal of Cross-Cultural Psychology, 26,* 153–168.

Lin, P. J., Schwanenflugel, P. J., & Wisenbaker, J. M. (1990). Category typicality, cultural familiarity, and the development of category knowledge. *Developmental Psychology, 26,* 805–813.

Marshall, C. E., & Parr, W. V. (1996). New Zealand norms for a subset of Battig and Montague's (1969) categories. *New Zealand Journal of Psychology, 25,* 24–29.

McEvoy, C. L., & Nelson, D. L. (1982). Category name and instance norms for 106 categories of various sizes. *American Journal of Psychology, 95,* 581–634.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105–142). New York: Academic Press.

Park, D. C., Nisbett, R., & Hedden, T. (1999). Aging, culture, and cognition. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 54,* P75–P84.

Pascaul, J. L., & Musitu, G. O. (1980). Normas categoriales [Categorical norms]. *Psicologica, 1,* 157–174.

Perfetti, C. A., Zhang, S., & Berent, I. (1992). Reading in English and Chinese: Evidence for a "universal" phonological principle. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 227–248). Amsterdam: North-Holland.

Pinto, A. C. (1992). Medidas de categorizacao: Frequencia de producao e de tipicidade [Category norms: Production, frequency, and typicality measures]. *Jornal de Psicologia, 10,* 10–15.

Piotrowski, M. (1998). *NLP-supported full-text retrieval* (CLUE Technical Reports). Erlangen, Germany: University Erlangen-Nuremberg.

Pollard, D. (2001). Distances and affinities between measures. In D. Pollard (Ed.), *Asymptopia* (chaps. 3 & 4). Retrieved June 15, 2004, from http://www.stat.yale.edu/~pollard/Asymptopia/Metrics.pdf

Posnansky, C. J. (1978). Category norms for verbal items in 25 categories for children in Grades 2–6. *Behavior Research Methods & Instrumentation, 10,* 819–832.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology, 4,* 328–350.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Erlbaum.

Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming scale useage heterogeneity: A Bayesian hierarchical approach. *Journal of the American Statistical Association, 96,* 20–31.

Salthouse, T. A. (1993). Speed and knowledge as determinants of adult age differences in verbal tasks. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 48,* P29–P36.

Shapiro, S. I., & Palermo, D. S. (1970). Conceptual organization and class membership: Normative data for representatives of 100 categories. *Psychonomic Monograph Supplements, 3,* 107–127.

Skrondal, A., & Rabe-Hesketh, S. (2003). Multilevel logistic regression for polytomous data and rankings. *Psychometrika, 68,* 267–287.

Smith, J. J. (1993). Using ANTHROPAC 3.5 and a spreadsheet to compute a free list salience index. *Cultural Anthropology Methods, 5*(3), 1–3.

Squire, L. R., & Kandel, E. R. (1999). *Memory: From mind to molecules.* New York: Scientific American Library.

Zelterman, D. (1987). Goodness-of-fit tests for large, sparse multinomial distributions. *Journal of the American Statistical Association, 82,* 624–629.

(*Appendix follows*)

Appendix

105 Categories and Degrees of Equivalence: All Culture-by-Age Groups

Category no.	Category description		American: young vs. old			Chinese: young vs. old			Young: American vs. Chinese			Old: American vs. Chinese		
	Americans	Chinese	HA	JS	Rank	HA	JS	Rank	HA	JS	Rank	HA	JS	Rank
1	Extinct animal	濒临灭绝的动物	.787	.844	88	.824	.863	66	.088	.351	104	.091	.348	102
2	Farm animal	牲畜	.956	.958	11	.985	.986	3	.895	.908	7	.905	.916	4
3	Four-footed animal	四足动物	.951	.958	13	.964	.971	7	.653	.704	31	.726	.783	18
4	Wild animal	野生动物	.937	.952	19	.907	.926	31	.776	.829	14	.809	.849	11
5	Major appliance	主要家用电器	.940	.953	16	.926	.932	22	.573	.667	46	.618	.711	31
6	Small kitchen appliance	厨房小电器	.837	.877	71	.824	.872	67	.238	.455	94	.202	.426	92
7	Auto parts	汽车部件	.894	.916	45	.815	.862	72	.608	.707	39	.559	.678	39
8	Bathroom fixture	浴室设备	.887	.909	49	.812	.858	74	.190	.418	101	.128	.382	99
9	Alcoholic beverage	酒精饮料	.899	.918	43	.882	.906	45	.606	.711	40	.543	.662	42
10	Non-alcoholic beverage	非酒精饮料	.878	.898	53	.781	.834	86	.376	.527	77	.465	.594	61
11	Bicycle parts	自行车部件	.930	.944	25	.868	.896	52	.761	.808	16	.665	.744	23
12	Bird	鸟类	.836	.872	73	.934	.950	17	.468	.602	64	.423	.570	69
13	Part of a boat	船的部分	.863	.893	62	.805	.847	77	.608	.709	38	.548	.665	41
14	Type of reference book	参考书种类	.814	.858	80	.526	.660	104	.265	.458	91	.328	.492	80
15	Type of bread	面包种类	.890	.906	48	.816	.857	71	.229	.424	97	.290	.477	83
16	Type of building	建筑物种类	.815	.857	79	.731	.800	92	.340	.499	86	.340	.509	78
17	Part of a building	建筑物的一部分	.871	.902	58	.791	.846	83	.676	.759	27	.592	.706	33
18	Building for religious services	宗教场所	.939	.953	17	.738	.801	91	.753	.827	18	.512	.641	47
19	Building material	建筑材料	.880	.911	51	.894	.915	38	.622	.697	36	.515	.618	46
20	Burial place	墓地	.784	.830	89	.677	.773	101	.163	.397	102	.095	.353	100
21	Type of candy	糖果	.743	.798	97	.835	.872	62	.398	.555	75	.278	.483	85
22	Chemical element	化学元素	.868	.899	59	.871	.897	50	.830	.863	12	.774	.831	14
23	Circus act	马戏节目	.777	.833	91	.631	.725	103	.270	.455	90	.177	.413	94
24	City	城市 (列举城市名称)	.790	.841	86	.961	.969	9	.200	.403	99	.081	.340	104
25	Type of cloth	衣料	.878	.903	53	.849	.878	59	.565	.683	48	.590	.700	34
26	Article of clothing	服装	.826	.858	75	.826	.862	65	.572	.672	47	.476	.612	57
27	Color	颜色	.954	.959	12	.976	.978	5	.904	.911	5	.927	.935	3
28	Means of communication	通讯工具	.825	.862	76	.886	.913	44	.541	.647	53	.568	.672	37
29	Compass direction	方向	.944	.947	15	.990	.991	1	.583	.630	44	.776	.795	13
30	Cosmetic	化妆品	.861	.879	63	.790	.838	84	.545	.663	52	.575	.682	35
31	Country	国家	.912	.932	35	.887	.894	43	.398	.532	74	.415	.544	71
32	Crime	罪行	.886	.915	50	.860	.893	54	.483	.595	61	.463	.597	62
33	Dairy product	奶制品	.985	.987	1	.897	.916	35	.297	.498	88	.338	.493	79
34	Type of dance	舞蹈种类	.701	.754	100	.847	.878	60	.586	.681	43	.370	.518	76
35	Disease	疾病	.671	.733	102	.718	.782	95	.499	.615	60	.411	.561	72
36	Breed of dog	狗的品种	.817	.851	78	.844	.884	61	.128	.373	103	.092	.361	101
37	Emotion	情感	.823	.852	77	.694	.765	99	.345	.513	84	.445	.590	65
38	Energy	能量	.729	.787	98	.891	.919	39	.678	.756	26	.664	.742	24
39	Source of energy	能量来源	.770	.825	94	.798	.841	80	.667	.752	30	.671	.755	22
40	Piece of farm equipment	农具	.877	.905	56	.881	.908	47	.540	.622	54	.567	.646	38
41	Piece of fire fighting equipment	灭火器具	.867	.900	61	.753	.814	88	.243	.452	93	.242	.449	91
42	Item of hiking equipment	徒步旅行的工具	.900	.923	42	.764	.823	87	.539	.654	55	.525	.655	44
43	Kind of explosive	爆炸的种类	.804	.846	84	.806	.857	76	.422	.570	70	.383	.546	75
44	Part of a face	脸部器官	.972	.975	6	.988	.990	2	.901	.911	6	.880	.902	6
45	Fish	鱼类	.907	.929	38	.888	.910	42	.356	.522	80	.258	.459	87
46	Tropical fish	热带鱼	.669	.751	104	.417	.579	105	.293	.500	89	.146	.404	97
47	Type of flower	花	.858	.886	66	.808	.850	75	.463	.592	67	.271	.447	86
48	Folk arts	民间艺术	.670	.756	103	.751	.818	89	.198	.421	100	.154	.394	95
49	Substance for flavoring food	调料	.911	.933	37	.888	.904	41	.563	.653	49	.506	.611	49
50	Type of footwear	鞋类	.859	.885	65	.957	.963	10	.349	.508	81	.249	.448	89
51	Fruit	水果	.962	.966	9	.972	.977	6	.839	.868	11	.866	.890	8
52	Type of fuel	燃料	.808	.853	81	.897	.921	34	.394	.555	76	.366	.520	77
53	Article of furniture	家具	.901	.914	41	.870	.892	51	.600	.701	41	.744	.803	16
54	Herb	草本植物	.837	.877	71	.659	.747	102	.080	.358	105	.050	.337	105
55	Herbal medicine	草药	.555	.682	105	.706	.781	98	.201	.427	98	.085	.365	103
56	Part of the human body	人体部分	.913	.927	33	.910	.922	29	.794	.831	13	.761	.796	15
57	Type of human dwelling	人类住所	.911	.931	36	.818	.863	70	.349	.509	82	.310	.484	81
58	Insect	昆虫	.926	.941	30	.930	.945	21	.747	.809	20	.680	.758	21
59	Piece of jewelry	珠宝	.926	.938	29	.920	.934	24	.247	.423	92	.138	.365	98
60	Kitchen utensil	厨房用具	.835	.864	74	.876	.904	48	.330	.481	87	.245	.449	90
61	Mathematical operation	数学运算	.957	.968	10	.944	.952	14	.888	.902	8	.941	.956	2

Appendix (*continued*)

Category no.	Category description		American: young vs. old			Chinese: young vs. old			Young: American vs. Chinese			Old: American vs. Chinese		
	Americans	Chinese	HA	JS	Rank	HA	JS	Rank	HA	JS	Rank	HA	JS	Rank
62	Major type of meat	肉类	.891	.907	46	.983	.986	4	.702	.773	23	.780	.831	12
63	Medical specialty	医科种类	.800	.839	85	.934	.944	17	.418	.567	71	.458	.599	63
64	Type of metal	金属	.904	.917	39	.896	.909	36	.853	.877	9	.828	.854	9
65	Military title	军衔	.932	.946	24	.850	.868	58	.501	.628	59	.476	.608	56
66	Type of motion	动作	.750	.817	96	.789	.843	85	.528	.658	57	.492	.627	53
67	Type of music	音乐类型	.771	.816	93	.691	.757	100	.464	.602	66	.469	.600	60
68	Musical instrument	乐器	.962	.969	8	.881	.900	46	.731	.794	22	.637	.730	30
69	Mythical being	神话人物	.701	.778	99	.714	.787	97	.235	.459	96	.147	.400	96
70	Natural earth formation	天然地貌	.878	.909	52	.715	.791	96	.477	.593	63	.408	.552	73
71	Non-relative relations	非亲属关系	.775	.823	92	.857	.893	55	.621	.716	37	.436	.592	68
72	Type of nut	坚果种类	.890	.905	47	.925	.941	23	.686	.747	25	.646	.736	28
73	Occupation or profession	职业	.806	.856	82	.857	.888	55	.643	.715	32	.516	.637	45
74	Internal organs	体内器官	.974	.979	5	.949	.955	12	.934	.945	2	.942	.955	1
75	Precious stone	宝石	.930	.939	26	.934	.949	16	.773	.828	15	.658	.720	26
76	Type of reading material	读物种类	.930	.949	27	.796	.849	82	.690	.759	24	.661	.744	25
77	Type of relative	亲戚关系	.904	.913	40	.833	.869	64	.375	.551	79	.438	.585	67
78	Religious object	宗教经典	.788	.834	87	.743	.808	90	.459	.600	68	.257	.458	88
79	Rodent	啮齿类动物	.915	.929	32	.909	.930	30	.344	.499	85	.285	.462	84
80	Room in house	房间	.965	.971	7	.906	.932	32	.675	.754	28	.649	.732	27
81	Member of royalty	皇室成员	.982	.986	3	.905	.922	33	.554	.680	51	.504	.643	50
82	Type of science	科学	.784	.819	90	.726	.771	93	.748	.799	19	.549	.660	40
83	Season of the year	季节	.976	.979	4	.933	.949	19	.907	.920	4	.880	.910	6
84	Type of shape	形状	.840	.862	70	.931	.946	20	.736	.793	21	.499	.604	52
85	Type of ship	船	.841	.876	69	.873	.897	49	.554	.664	50	.530	.636	43
86	Type of singing voice	嗓音的音域类别	.935	.950	22	.853	.893	57	.626	.703	34	.684	.751	20
87	Snake	蛇	.852	.880	67	.911	.925	27	.532	.632	56	.449	.576	64
88	Social science	社会科学	.696	.773	101	.800	.844	78	.626	.703	33	.485	.609	54
89	Sport	运动	.917	.931	31	.823	.861	68	.376	.526	78	.305	.470	82
90	Type of storm	风暴类型	.947	.957	14	.833	.875	63	.449	.592	69	.502	.624	51
91	Surgical instrument	手术器具	.805	.851	83	.796	.831	81	.595	.710	42	.472	.606	59
92	Unit of time	时间单位	.983	.986	2	.911	.918	28	.962	.970	1	.817	.853	10
93	Time of day	一天的时间	.936	.950	20	.917	.935	25	.754	.815	17	.739	.809	17
94	Carpenter's tool	木工用具	.861	.887	64	.865	.892	53	.622	.688	35	.642	.704	29
95	Gardening tool	园艺工具	.896	.916	44	.798	.848	79	.506	.612	58	.512	.618	48
96	Toy	玩具	.753	.809	95	.725	.782	94	.236	.443	95	.194	.413	93
97	Tree	树	.876	.898	57	.916	.934	26	.483	.597	61	.443	.559	66
98	Part of a tree	树的部分	.927	.938	28	.961	.966	8	.928	.944	3	.887	.913	5
99	Vegetable	蔬菜	.932	.943	23	.946	.956	13	.467	.606	65	.395	.551	74
100	Type of vehicle	车辆	.878	.895	55	.889	.908	40	.348	.523	83	.485	.608	55
101	Type of waterway	水路	.938	.951	18	.815	.867	73	.671	.751	29	.608	.705	32
102	Weapon	武器	.868	.894	60	.896	.917	37	.575	.662	45	.571	.656	36
103	Weather phenomenon	天气现象	.936	.949	21	.950	.958	11	.412	.554	72	.474	.589	58
104	Kind of wood	木材	.842	.872	68	.818	.862	69	.404	.552	73	.419	.569	70
105	Writing implement	书写工具	.912	.922	34	.936	.945	15	.841	.882	10	.719	.782	19

Note. Ranks refer to degree of similarity based on HA values for each between–culture-by-age-group comparison. HA = Hellinger Affinity; JS = Jensen–Shannon entropy.

Received July 22, 2003
Revision received December 3, 2003
Accepted March 26, 2004 ■