Revisiting Tversky's Diagnosticity Principle

Ellen R. K. Evers*

Tilburg University

Daniël Lakens

Eindhoven University of Technology

Correspondence:

Ellen R. K. Evers

Tilburg University

Department of Social Psychology & TIBER

PO Box 90153, 5000 LE, Tilburg, The Netherlands

E.r.k.evers@tilburguniversity.edu

Introduction

In 1977 Amos Tversky published a paper that critiqued geometric models of similarity, and proposed an alternative approach known as the contrast model. Since its publication, the article has been cited 5343 times (Google Scholar). Tversky's contrast model describes several principles people rely on when they judge the similarity of stimuli. The diagnosticity principle, the focus of the current replication proposal, is one of these principles.

The diagnosticity principle states that features of stimuli that are relevant for categorization will receive relatively more weight when people perform similarity judgments than non-diagnostic features. In other words, when people judge similarity, the features of the stimuli that these judgments are based on are highly dependent on the set of objects under consideration. Tversky illustrates this idea with the following example: "the feature "real" has no diagnostic value in a set of actual animals since it shared by all actual animals and hence cannot be used to classify them. This feature, however, acquires considerable diagnostic value if the object set is extended to include legendary animals, such as a centaur, a mermaid, or a phoenix" (Tversky, 1977, pp.342).

Importance

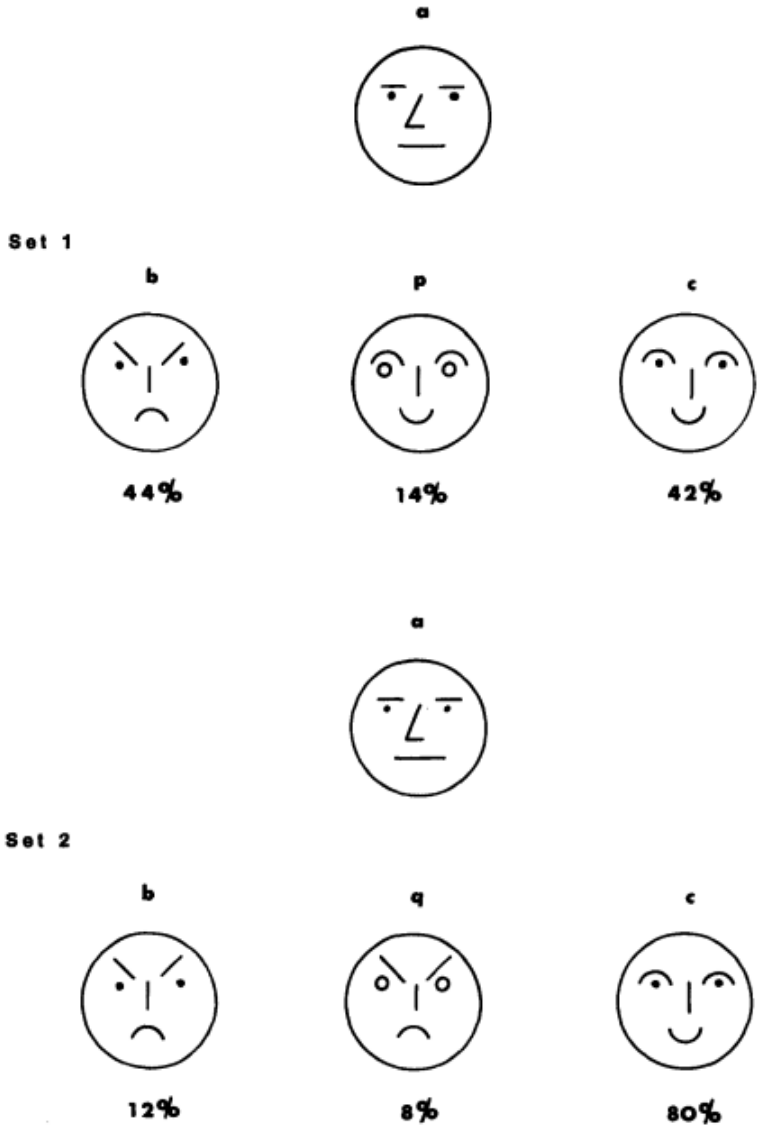
As mentioned in the introduction, Tversky model of similarity was a major contribution to the field of psychology. Even though follow-up research has expanded upon this model and has made several adjustments (see for example Genter & Markman, 1997), the main premises of the model are still accepted today, as indicated by the large number of citations. In other words, Tversky's paper led to a fundamental change in how scientists think about similarity judgments (for a discussion, see Goldstone &Son, 2005).The diagnosticy principle is especially interesting for psychologists, because it revealed how geometric models of similarity failed to take into

account that human cognition is inherently context-dependent. It is the context that determines how different features of stimuli are weighed in similarity judgments. Tversky's diagnosticity principle contributed to what is currently understood to be a basic principle of similarity judgments. This is also revealed in the number of citations that explicitly mention the words "diagnosticity" or "diagnostic" (259 of the papers that cite Tversky 1977 mention the word "diagnosticity" and a total of 836 articles mention the word "diagnostic"). It is therefore remarkable that only a modest number of studies have tried to (conceptually) replicate the diagnosticity effect. Furthermore, whereas the original studies have yielded clear results, follow up studies do not allow unequivocal conclusions, because all these replication attempts either test the diagnosticity effect in a different way (which makes effect-sizes incomparable with the original studies), or suffer from methodological or data-analytical problems. Therefore, there is a general imbalance between the theoretical importance of the diagnosticity effect, and the amount of empirical support for its existence. We aim to resolve this discrepancy through the proposed replications.

Studies on diagnosticity in Tversky 1977

Tversky reports two different studies that provide support for the diagnosticity effect. The first study reported in Tversky 1977 uses a straightforward paradigm. First, participants were shown a group of four faces (see Figure 1, below).

Figure 1: Stimuli used in Experiment 1, Tversky, 1977

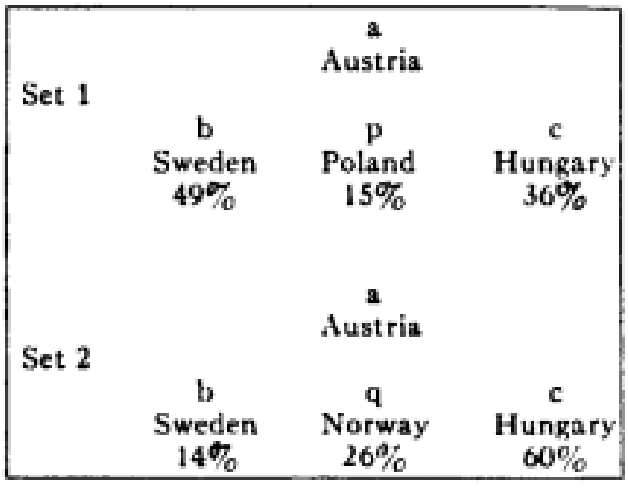


This group of faces always consisted of a neutral (a, the target stimulus on the top), frowning (b), &smiling (c) face. For half of the participants the fourth face was p (smiling), for the other half it was q (frowning). They were subsequently asked to split the four faces into two groups of two faces. As expected, most participants grouped c&p (smiling) and a&b (non-smiling) in the first condition, but b&q (frowning) and a&c (non-frowning) in the latter. According to the diagnosticity principle, smiling vs. non-smiling thus has a greater diagnostic

value in the first set of faces, but frowning vs. non-frowning is more diagnostic in the second set of faces. Subsequently, a different group of participants (N = 50) was asked to pick the face that most resembled the target face (a) from a group of b, p[q], and c. Mirroring the pattern of asking participants to group the faces, participants were more likely to pick the frowning face as most similar to the neutral target face when the other two faces were smiling (condition 1), as compared to when two of the three faces were frowning (condition 2, for the choice proportions, see figure 1). The reverse was true for the smiling face, which was less likely to be picked as similar to face a in condition 1 as compared to condition 2.

Study 2 on the diagnosticity principle in Tversky 1977 is a summary of a study described in more detail as Experiment 4 in Tversky & Gati, 1978. It is an extension of the previous study and used semantic stimuli (See figure 2, below)

Figure 2: Stimuli used in Experiment 2, Tversky, 1977 / Experiment 4 Tversky & Gati, 1978



First, an independent group of participants was asked to cluster 4 countries. Three countries were the same in each condition (a, b, & c) but the fourth varied (p / q). Instead of a single trial experiment (as in Experiment 1), 20 (x 2) sets of countries were created.

After this categorization task, a new group of participants (N = 33) was asked to make similarity judgments (as Experiment 1). It was expected that similarity judgments would follow the same pattern as the categorizations made by the previous group of participants, and this pattern of results was indeed observed. Whereas in Experiment 1 the diagnosticity of features of novel stimuli (schematic faces) emerges within the task, in Experiment 2 diagnosticity is dependent on participants' previous knowledge about the stimuli (characteristics of countries). Other than that, the conclusions from Experiment 2 closely resembled those of Experiment 1. For the full results of Experiment 2, see appendix A.

Replications

As already mentioned in the introduction, even though Tversky uses a fairly straightforward experimental method, and despite the theoretical influence of the article, there is a remarkable absence of replication attempts. Even more striking, all attempts we were able to find could either not be compared directly with the original studies by Tversky, were flawed, or had no strong evidential value for a diagnosticity effect.

Direct replications

A literature search revealed one published close replication, in which participants from two different cultures (China and Australia) performed the original experiment (Zhou, Fu, Hayward, Locke, & Peillicano, 2005). The results of this experiment are rather difficult to interpret. First of all, the average number of choices are not given in the text but must be estimated from a graph portraying the percentages. Because of a mistake in the description of the

number of participants (the reported sample size does not match the total of Chinese and Australian participants) it is impossible to calculate the number of participants in each condition, and therefore how many participants made which choice. A further problem with these data is that the crucial test for the diagnosticity-effect (comparing choices in set 2 with choices in set 1) is not reported. Instead, the authors test whether the choice proportions by the Chinese and Australian participants were different from each other, and different from Tversky's participants. In addition to the difficulties of interpreting the data, the Australian conditions consisted of (approximately) 38 participants divided over 2 conditions, giving the experiment only sufficient power if the effect-size one is interested in is extremely large. It is difficult to draw strong conclusions from this replication study.

The only other direct replication we know of is an unpublished study by one of the authors of this proposal (Lakens, unpublished). This replication was conducted in the classroom (similar to the original studies by Tversky). The sample size was relatively small (n_1 = 48, n_2 = 41). Even though the study was probably underpowered, the results were still insightful since the null-effect was a trend in opposite direction of the hypothesized (and previously found) effect. Taking Tversky's results as the real expected frequencies, the chance of finding a trend in opposite direction is very small. Given the lack of an effect in the replication study, we should seriously consider the possibility that Tversky's (1977) effect-size is an overestimation of the true effect.

Conceptual replications

A conceptual replication (Goldstone, Medin, & Halberstadt, 1997, Experiment 2A and 2B) used schematic faces similar to those used by Tversky (1977) with one crucial difference; the distracter face and one of the other faces shared a feature that did not match the comparison

face on some trials (shared mismatch) and shared a feature that was also possessed by the comparison face on other trials (shared match).Furthermore, the authors attempted to decrease the number choices for the distracter option by participants. The reason for this is that in Tversky's original work it is unclear which part of the effect of diagnosticity could be explained by a mere substitution effect.[1] Even though Goldstone, Medin,& Halberstadt do find a diagnosticity effect, the size of this effect is much smaller than the original effect found by Tversky (averaging 35-40% in Tversky's original studies, and only 4.5 – 5% in the studies by Goldstone and colleagues). Since the authors only used a very small sample for the initial categorization task (n = 10, n=18 respectively), it is unclear whether this small effect is results from ambiguous stimuli, or whether it means that a large part of the effect observed by Tversky was not a diagnosticity effect, but a substitution effect. Another possible explanation for the difference in effect sizes is the fact that in these studies participants were asked to judge the similarity in a total of 160 trials per person. This makes it possible that participants adopted a context-independent system of preferences.

The other conceptual replication, conducted by Medin and Kroll (unpublished dataset, reported in Medin, Goldstone, & Markman, 1995), used geometrical shapes, and attempted to examine whether the diagnosticity effect occurred above and beyond a substitution effect .They also observe an effect of diagnosticity, but again only a fairly small effect (55% of the time the diagnostic option is chosen over the non-diagnostic one). Furthermore, by asking participants to rank the options in degree of similarity, they adjusted the method, which might cause participants to use a different strategy for the selection for the most similar option.

In other words, both these studies are currently cited as support for the diagnosticity principle, but their results cannot be compared to Tversky's original findings. The choice

percentages do support the existence of a diagnosticity effect, but the effect seems much smaller than in Tversky's original study. On average, the authors find that the diagnostic option is chosen over the non-diagnostic option 51-55% of the time.

Conclusion introduction

To summarize, evidence for the diagnosticity effect is mixed, and limited. We know of five close replication attempts. Of these five experiments, one is interpretable but likely underpowered (Lakens, unpublished).Two experiments do not report the data needed to test whether a diagnosticity effect was observed, and if so what the effect-size was (Zhou et al., 2005).The final two experiments (Goldstone et al., 1997; Medin & Kroll in Medin, et al., 1995) do replicate a diagnosticity effect, but this effect is much smaller than the original diagnosticity effect found by Tversky (1977).These differences in effect size could be a consequence of the differences in methodology, but it is also possible that Tversky's original findings are by large the result of a substitution effect misinterpreted as diagnosticity. Because of the importance of the diagnosticity effect, and the limited empirical evidence for it, we believe that the original work by Tversky (1977) is an excellent candidate for replication. Therefore, we plan to replicate both Study 1 and Study 2 reported in Tversky, 1977. However, since the original findings are possibly (partially) the result of substitution effects we will expand the original studies in such a way that we first replicate Tversky's original method, and subsequently add questions designed to tease the substitution effect apart from the diagnosticity effect.

Proposed Replication

Participants

We plan to replicate both studies using a Dutch student sample, an American sample on mTurk, and an Israeli sample through an internet questionnaire. Although in the original paper,

no description of the participants is provided in Tversky (1977), it is stated in Tversky and Gati (1978) that participants were undergraduate students majoring in the social sciences from the Hebrew University in Jerusalem and the Ben-Gurion University in Beer-Sheba. Theoretically, no moderating effect should be expected for age or gender. Based on Zhou et al (2005) cultural differences between Western and Eastern participants might moderate the effect. There is no reason to expect differences between cultures to influence our replication studies since we will only recruit students from 'Western' countries (i.e., Israel, The Netherlands, The United States).

Method

We plan to replicate the original method as closely as possible. For the Dutch participants, we will use paper and pencil questionnaires but we will have to conduct the experiment for Israeli and US participants electronically. Theoretically, this difference should not matter. Study 1 can be replicated directly. Study two requires some of the countries used as stimuli to be replaced, since these countries no longer exist (i.e., Czechoslovakia, Yugoslavia, U.S.S.R. and West-Germany). We will replace those countries by new countries (e.g., Czech Republic, Serbia, Russia, Germany). Before running the replication studies we will perform a pilot test that repeats the classification-task to check whether the classifications with the new countries is identical to the classifications in the original paper. In cases of multiple replacement-options, we will test all and choose the replacement-country that has the largest effect on classification. In the case of no of the new trial version showing the effect on classification, it will be excluded from the replication. After the direct replication questions (e.g., please indicate the face [country] most similar to the face [country] portrayed at the top of the page), we will ask the participants to rank all three options in order of most similar to least similar. This gives us the

opportunity to not only compare the effect size found with the original findings of Tversky, but also allows us to interpret the effect-size of diagnosticity excluding possible substitution effects.

Sample size

Even though replications are essential for a well-functioning science, it is remarkably unclear what constitutes a successful or unsuccessful replication. It is possible that a replication attempt fails to reject the null, while still finding an effect-size close to the original experiment. It is also possible that a replication attempt rejects the null but finds an extremely lower effect-size. We therefore believe that it is important to not only have a sample-size that is determined to be big enough to refute the null (general power analysis), but also to have a sample size that is powerful enough to be able to refute the alternative null (that of the effect size being "x"). Therefore we plan to follow Simonsohn's (2013) advice of using the original N * 2.5. We plan to analyze the data both using standard null-hypothesis-testing, as well as Bayesian statistics.

Analysis

Data Screening. In Experiment 1, participants perform a single judgment that does not provide any justification to exclude participants, and therefore, following the original study, we will analyze all data without excluding participants. In Experiment 2, participants provide 20 judgments by choosing between three randomized response options. This allows us to examine practices such as straight lining (e.g., choosing the left answer possibility throughout the experiment). In the original study, no participants were excluded from the analysis for any reason. We will not exclude participants, but will report the percentage of 'straightliners' (which we define as choosing responses on the same location at least 19 out of 20 times), which should help to interpret the observed effect size.

Data Analysis First, we will analyze the results is exactly the same ways how they were originally analyzed by Tversky (1977). This means that for Experiment 1, choices for the distracter options p[q]will be excluded from the analysis and that choice proportions for the other two faces will be analyzed using a chi-squared test. In addition, the effect size Cramer's *V* (including the 95% confidence interval around the effect size estimate) will be calculated and reported. For Experiment 2, we will calculate the proportion of choices for country a when it is the diagnostic option, minus the proportion of choices when it is not the diagnostic option (p_{ad}– p_{an}). We will do the same for country b ($p_{bd} - p_{bn}$). If there is no effect of diagnosticity, both these difference-scores should average around 0. However, if there is a diagnosticity effect, the difference scores should be higher than 0.Using a one sample *t*-test, these scores can be compared to 0 as an overall test for the effect, and we will calculate Cohen's *d* and the 95% confidence intervals around this effect size estimate. Furthermore, a chi-square test will be conducted and reported for each separate trial (every single country). In addition to the null-hypothesis significance tests, we will perform a Poisson Exponential ANOVA (following Kruschke, 2010), the Bayesian equivalent of the Chi-square test, and a one-sample *t*-test (following Kruschke, 2013), assuming an uniform prior, and report the Bayes factor (following Rouder, Speckman, Sun, Morey, & Iverson), as well as the 95% highest density interval.

The second analysis will follow that of Medin, Goldstone, & Markman (1995) who used similarity rankings. In this analysis we will not compared the proportion of choices for the diagnostic option with the proportion of choices for the non-diagnostic option, but instead we will look at the proportion of participants ranking the diagnostic option over the non-diagnostic one. This way, we control for the substitution effect. These proportions based on similarity rankings can then again be analyzed in the same way; a chi-square for Experiment 1, a one

sample *t*-test for Experiment 2, and finally individual chi-squares per trial to estimate the overall effect-size (including the calculation of effect sizes and 95% confidence intervals around the effect size estimate). Using these two methods does not only allow us to investigate how well Tversky's original findings replicate, but by using the 2nd analysis we are also able to estimate the effect-size of diagnosticity while excluding effects due to substitution. This also allows us to examine whether Tversky's original findings were overinflated estimates due to substitutability, or whether subsequent (much lower) effect sizes were the result of a different method of eliciting judgments of similarity.

Finally, since the diagnosticity effect is theorized to be the result of categorization, we will correlate the item level scores on categorization (part 1 of the experiments) with the average proportions given in the similarity judgment task. Even though this was not done in the original (or the follow-up) studies, we believe this is a valuable addition, since such a correlation would be predicted based on the theory. Finding such a correlation would provide additional support for the existence of the diagnosticity effect, as well as for the underlying process. Failure to find such a correlation would indicate that another process may be the underlying cause of the effects on similarity.

Interpretation of the Data Analysis. Our main interest in these replication studies is to increase the confidence in the diagnosticity principle proposed by Tversky (1977). Because Tversky was not able to distinguish the diagnosticity effect from a mere substitution effect in his studies, and subsequent studies that did correct for substitution effects are difficult to compare with the original studies. We will first see whether we can replicate the effects Tversky found. Then we will control for possible substitution effects, any remaining reliable effect (in other words, whenever the 95% confidence interval of the effect sizes does not include 0, or when the

95% HDI excludes 0) will be interpreted as support for the diagnosticity principle. If we do not observe a reliable difference, we can interpret the Bayes factor in terms of the strength of evidence for the null-hypothesis.

References

Goldstone, R. L., Medin, D. L. & Halberstadt, J. (1997). Similarity in context. *Memory &Cognition, 25*, 237-255.

Goldstone, R. L, & Son, J. (2005). Similarity. In K. Holyoak & R. Morrison (Eds.). *Cambridge Handbook of Thinking and Reasoning, 13-35*. Cambridge: Cambridge University Press.

Kruschke, J. (2010). *Doing Bayesian Data Analysis: A Tutorial Introduction with R*. Academic Press.

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General, 142*, 573-603.

Medin, D. L., Goldstone, R. L., &Markman, A. B. (1995). Comparison and choice: Relations between similarity processes and decision processes. *Psychonomic Bulletin & Review, 2*,1-19.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review, 16*, 225–237. doi:10.3758/PBR.16.2.225

Simonsohn, U. (2013). Evaluating Replication Results. Retrieved May 6th, 2013 from http://dx.doi.org/10.2139/ssrn.2259879

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-352.

Tversky, A., & Gati, I. (1978). Studies of similarity. *Cognition and categorization, 1*, 79-98.

Zhou, G., Fu, X., Hayward, W., Locke, V., & Peillicano, E. (2005). Cultural Difference in the Application of the Diagnosticity Principle to Schematic Faces. *Journal of Cognition and Culture, 5*, 240-24

Footnotes

1. The substitution effect occurs when two (or more) options in a choice-set have a positive cross-elasticity. For example; when choosing between a chocolate bar and an apple, people who care about health will probably choose the apple over the chocolate bar. However, when adding an orange to the choice set, health-conscious people may either choose the apple or the orange, so the proportion of people choosing the apple may go down, not because people like the apple less, but purely because they would have chosen the apple for its health aspect and now have two options sufficing that aspect. Similarly, people who are asked to indicate which item is most similar to a banana out of an apple and the sun may either focus on the aspect "fruit" or the aspect "color". Addition of a yellow submarine would therefore decrease choice for the sun, purely because now there are two yellow options, and addition of an orange may decrease choices for the apple, because there are two fruits. Such findings would resemble those predicted by a diagnosticity effect but would clearly not be the result of diagnosticity as theoretically formulated by Tversky (1977)

Appendix A

Below a reproduction of the table of results from Experiment 2 in Tversky 1977. Participants were asked which country was most similar to country *a* out of a set consisting either of country *b*, *c* and *q*, or *b*, *c* and *p*. Reported in the three right-hand columns are respectively; "percentage of choices for *a* being most similar to *b* when the set includes *p*, minus the percentage of choices for *a* being most similar to *b* when the set includes *q*" ($b(p)$-$b(q)$); "percentage choices for *c* when the set includes *q*, minus the percentage of choices for *c* when the set includes *p*" ($c(q)$-$c(p)$); and finally, the difference between the proportion pairing *a* with *b* over *c* when the set includes *p*, minus the difference between the proportion pairing *a* with *b* over *c* when the set includes *q* ($D(p,q) = a_p(b,c) - a_q(b,c)$. The choice-proportions were not reported for this experiment and cannot be reconstructed with this information.

Table 3.4
Classification and Similarity Data for the Test of the Diagnosticity Hypothesis

	a	*b*	*c*	*q*	*p*	$b(p) - b(q)$	$c(q) - c(p)$	$D(p, q)$
1	U.S.S.R.	Poland	China	Hungary	India	6.1	24.2	66.7
2	England	Iceland	Belgium	Madagascar	Switzerland	10.4	−7.5	68.8
3	Bulgaria	Czechoslovakia	Yugoslavia	Poland	Greece	13.7	19.2	56.6
4	U.S.A.	Brazil	Japan	Argentina	China	11.2	30.2	78.3
5	Cyprus	Greece	Crete	Turkey	Malta	9.1	−6.1	63.2
6	Sweden	Finland	Holland	Iceland	Switzerland	6.5	6.9	44.1
7	Israel	England	Iran	France	Syria	13.3	8.0	87.5
8	Austria	Sweden	Hungary	Norway	Poland	3.0	15.2	60.0
9	Iran	Turkey	Kuwait	Pakistan	Iraq	−6.1	0.0	58.9
10	Japan	China	W. Germany	N. Korea	U.S.A.	24.2	6.1	66.9
11	Uganda	Libya	Zaire	Algeria	Angola	23.0	−1.0	48.8
12	England	France	Australia	Italy	New Zealand	36.4	15.2	73.3
13	Venezuela	Colombia	Iran	Brazil	Kuwait	0.3	31.5	60.7
14	Yugoslavia	Hungary	Greece	Poland	Turkey	9.1	9.1	76.8
15	Libya	Algeria	Syria	Tunis	Jordan	3.0	24.2	73.2
16	China	U.S.S.R.	India	U.S.A.	Indonesia	30.3	−3.0	42.2
17	France	W. Germany	Italy	England	Spain	−12.1	30.3	74.6
18	Cuba	Haiti	N. Korea	Jamaica	Albania	−9.1	0.0	35.9
19	Luxembourg	Belgium	Monaco	Holland	San Marino	30.3	6.1	52.2
20	Yugoslavia	Czechoslovakia	Austria	Poland	France	3.0	24.2	39.6