Appendix A
Methodology of the State Department Study

Administration

Invitations to participate in the study were mailed, together with a questionnaire, to a random sample of 300 foreign service officers stationed in Washington, to the 100 military officers in the current class at the National War College (now the National Defense University), and were distributed by two contacts to 50 members of the domestic affairs section of the Office of Management and Budget.[1] Cover letters described the project in general terms and promised a report of the findings if the respondent completed and returned a separate form (this was to be returned in a separate envelope to preserve anonymity of the questionnaire itself). The State Department mailing also included a letter of introduction from David Biltchik, a foreign service officer who was head of the Face-to-Face program established by the Carnegie Endowment and the American Foreign Service Association. Follow-up letters were mailed to FSOs and military officers about a week after the first contact.[2]

The author was identified in these letters as a doctoral student associated with Yale University and holding a fellowship grant from the National Institute of Mental Health through the Yale Psychological Study of Politics program.

Sampling and Response Bias

The results of the study may be somewhat sensitive, in several respects, to details of its conduct and execution. Specifically:

1. The use of mailed questionnaires rather than personal contacts probably reduced significantly the response rate at the State Department and National War College (to 42% and 49%, respectively, compared with 74% at OMB).[3]

2. Members of the OMB staff were experiencing a temporary lull following completion of the president's budget; this and the time required to complete the lengthy questionnaire (45 minutes to an hour) probably also contributed to lower response rates at the State Department and NWC. This hypothesis was supported by the tendency of response rates to decrease (to 30%) at the higher levels (FSO-1 and FSO-2) in the State Depart-

ment. As a result, one key variable, the tendency to adopt a traditionalist view of the origins of the cold war, is probably slightly understated since this variable is associated with rank (as well as with age).

3. The questionnaire did not inquire about sensitive current issues (e.g., the SALT talks, relations with the Middle East, Vietnam, Cambodia, India, Pakistan, Chile, Taiwan, and the People's Republic of China) in favor of issues that would serve the theoretical purposes of the study for which candid responses would be more easily obtained. Nevertheless some diplomats and military officers declined to participate on the grounds it would be inappropriate for them to answer *any* questions. This attitude, which I assume may include a component of fear of the potential disapproval of one's superiors, probably means that the degree of psychological subordination to American foreign policy is understated in the results obtained. Thus the variables associated with psychological transcendence—altruism, generosity, reluctance to use force—are probably slightly less characteristic of military officers and FSOs in general than they are among those who agreed to participate.

4. It also became clear, both directly and indirectly, that some men felt deeply threatened and were, as a result, antagonistic to the study. I could not tell whether this was a result of a personal aversion to quantitative methods, to psychological probing, a mistrust of my motives, or a fear that the results—if any significant findings were obtained—might undermine the credibility of the State Department, or was a combination of these and other feelings. But a plausible guess is that those who did not respond were somewhat more mistrustful and somewhat more antagonistic than those who did participate.[4] Thus these tendencies are probably understated, and the associated tendencies to use force and to be fearful of the Soviet Union also ought to be somewhat increased if one wishes to extrapolate to all FSOs and military officers.

5. It is part of the folk wisdom of survey research (although there is little hard evidence) that individuals with low self-esteem or larger numbers of neurotic symptoms are less likely to respond voluntarily to psychological questionnaires.[5] If so, then some adjustment would also be needed in these scores and in the reported tendency to advocate the use of force.

In sum, then, those who did respond are probably not fully representative of those who did not respond. The results obtained probably overstate slightly the degree of trust, psychological transcendence, self-esteem, and mental health in these groups as a whole. And they probably understate slightly the prevalence of strong traditional views of the origins of the cold war, the fear of the Soviet Union, the tendency to use force, and overstate slightly the degree of altruism and generosity in the sample compared with those who did not respond.

History: Changes During the Course of Research

Former Ambassador John Bartlow Martin once observed that national policy can be "overtaken by events."[6] The same is true of researchers: one potential problem is that changes in the international or domestic political climate may occur while the research is in progress and diminish the comparability of results obtained at the beginning and at the end of the research period. Fortunately no such problem seemed to emerge here: the time of the study—December, 1971, to early February, 1972, was a relatively quiet period in international relations. The study commenced following the end of the latest India-Pakistan war and was concluded prior to President Nixon's previously announced trip to Peking. The only major news story was the revelation of secret transcripts by columnist Jack Anderson which showed that the administration, while officially and publicly neutral during the India-Pakistan war, was in fact ordered by President Nixon in private to "tilt toward Pakistan." However revelations of official duplicity were nothing new in Washington, and it is unlikely that such newspaper stories significantly affected the study, especially so because nothing in the study dealt with highly sensitive issues then under active discussion.[7]

Instrument-Subject Interactions: Order Effects

One problem with questionnaires is that the order of presentation of questions may introduce a halo effect that colors answers to later items; or a subject may become fatigued or otherwise affected while working his

way through the questionnaire and this may affect his answers. The trick is to minimize the extent to which such effects generate *systematic* patterns of response across subjects that can be mistaken for true correlations. To deal with this problem, questionnaire items were grouped into six units on the basis of response format, page orders within units were varied systematically, and the units were assembled in random order and then randomly assigned to subjects. Each individual received a different order of presentation selected at random. Thus the probability was minimized that subjects who would give a biased response to a given order would receive that order. In addition, the order of adjective scales within the semantic differential ratings, although constant for each object, was varied randomly across the six objects.[8]

It is also a possibility that an individual might guess what is being investigated and alter his answer to affect what would be concluded about him. This seems unavoidable in the case of policy questions and for several of the perceptual scales; but for the personality scales items were intermixed from other scales with identical format, or "buffer items" were included.[9]

Instrument-Subject Interactions: Clarity

An important question about the validity of the policy attitude measures is whether the questions posed and the options provided were clear and meshed with the naturally occurring categories of the respondents. Two procedures were used to deal with this issue: first, an initial draft of the questionnaire was used in a pilot study with a small group of foreign policy specialists and graduate students in international relations, and the questionnaire was revised to eliminate any ambiguities that emerged. Second, a tally was kept for all questionnaires to indicate those questions which produced additional comments or qualifications by respondents. This tally showed that only two questions prompted qualifications by more than 2% of respondents. In one case the question about policies toward nonexpansionist Communist governments in underdeveloped countries provided two options, "oppose them" or "accept them." Five

FSOs who checked "accept them" felt it important to add the cautionary note "but do not encourage them."

The only significant validity problem occurred with attributions of Soviet intentions in the Middle East. To force respondents to divide between essentially a muddling-through incrementalist conception and a long-range expansionist conception, a middle-ground option of "muddling-through expansionism" was not included. Twenty-eight percent of diplomats wrote in this kind of additional option on their own and checked it. (Since some who may have preferred this option may not have written it in, the 28% is probably an underestimate and the other percentages probably are slightly overrepresented.) But the demonstrated willingness of respondents to alter options with which they did not feel comfortable provides evidence that other questions did align readily with their true attitudes.[10]

Scales—General Remarks

Most of the personality trait scales used were standard scales developed by previous researchers: the Leary Interpersonal Checklist measured dominance and hostility in interpersonal relations, competitiveness, and provided the adjective self-portraits in chapter 3. The Survey Research Center Trust in People scale (a modified version of the Rosenberg Misanthropy scale) assessed trust. The short form of the Maudsley (Eysenck) Personality Inventory measured neurotic symptoms and extraversion. Where reported in the literature these scales have test-retest reliabilities of about .70 and also display face validity.

The six objects assessed by the standard semantic differential technique (Myself, What I Would Like To Be, American Foreign Policy, Ideal American Foreign Policy, Soviet Foreign Policy, British Foreign Policy) were each rated on 20 adjective dimensions. A factor analysis (orthogonal rotation, varimax criterion) performed on the State Department random sample generated the factorial structure employed for all groups: the *evaluative* factor consisted of the adjective pairs good-bad, skillful-bungling, kind-cruel, honest-dishonest, friendly-menacing, and trusting-fearful (the

friendly-menacing scale for British and Soviet foreign policies were also separated as single variables for further analysis); an *activity-power* factor (interpreted as an ambition score when referring to the ideal self) consisted of the adjective pairs: active-passive, fast-slow, moving-still, strong-weak, hard-soft, tenacious-yielding, aggressive-defensive, leading-following, dominating-submitting, and resolute-irresolute: a *pragmatism* factor consisted of the adjective pairs pragmatic-ideological and realistic-naive; and a *stability* factor consisted of the single adjective pair stable-changeable.[11]

Separate factor analyses of self objects (Myself, What I Would Like to Be) compared with the four foreign policy objects in the FSO random sample showed factor structures virtually identical with the combined factor structure.[12] However, factor structures decomposed by object and by group showed substantial variability across these 18 separate factor structures—implying that the very general summary analysis used in the text loses considerable information by forcing subtle and complex thought into a Procustean bed. This suggests that significantly greater explanatory power of theories dealing with personality effects on cognition and perception might be obtained with a larger data base and a multimethod strategy which clarifies the reasons for these differing structures.[13]

Scales—Tendency to Advocate Force

It is important to emphasize that the five scenarios reported in chapter 4 were *not* a random sample of situations confronting policy makers. No conclusion is warranted that anyone scoring 60% on this scale will tend to advocate the use of force in 60% of all future foreign policy decisions. Instead, the scenarios were based on recent American policy and were constructed by an intuitive process to present situations in which the overt use of American force would be considered a relevant option.

It is also important to emphasize that the scenarios were standardized, specifically worded to stipulate certain beliefs on the part of the respondent, usually a belief that the use of force would be an effective policy tool. This may have had the consequence of increasing slightly the number of respondents who were favorable to force compared with their policy views in an actual decision-making situation.

The internal structure of the scale deserves brief attention in three respects: reliability, undimensionality, and the potential confounding of direction with intensity in the analysis of such a Likert scale. The inter-item correlations of the five items were modest, averaging about .40. In other words, "tendency to use force" is not a strong trait in which the direction of a man's judgment can always be predicted across situations. Many factors enter into a final decision. The practical effect of this problem is that the moderate reliability of this scale means that the overall variance which *could* be explained even by a perfect theory is substantially reduced (see the discussion below on correction for attenuation and explanatory power), making the explanatory power of the results obtained (relative to that which might have been explained with greater prior knowledge and a more sophisticated scale) greater than the R^2 statistics which are reported.

One subtle aspect of the tendency to use force is illustrated by the finding that it is rather consistently affected by the same personality traits across all five scenarios, except that a possible confrontation with the Soviet Union is not affected by competitiveness while using troops in the four other scenarios is consistently affected by variations in this trait. (In the text this difference is attributed to the fact that the other four scenarios deal with underdeveloped countries. But, in truth, the 5-item scale used in the study is too preliminary to be certain that this was the operative difference).

A final problem is that the scenarios were presented as 7-point Likert scales (ranging from "strongly advocate" to "strongly oppose"). Allowing such a weight for intensity to enter into a study of the relation between affect and decision, a common practice, potentially confounds those personality traits that affect *intensity* of advocacy with those that affect *direction*. To avoid this confounding, the variable was scored to take into account only the *direction* of the policy decision (% of cases), not its intensity. This is the more circumscribed analysis reported in the text.[14]

A Note on Models

The regression results reported are the best-fit equations resulting from the following general model:

$Y = a_0 + b_1X_1 + b_2$(NWC shift) + b_3(OMB shift) + b_4(NWC shift) X_1 + b_5 (OMB shift) X_1.

The term a_0 is a derived constant, the intercept; b_i is the coefficient of the ith term; X_1 is the independent variable (scaled from 0 to 10 except for competitiveness, which is 0 or 1), Y is the dependent variable (scaled from 0 to 10 except for the tendency to use force which goes from 0% to 100% and the desired force level which goes from ½ to 2½.) The additional interaction terms were included to test whether the National War College and Office of Management and Budget respondents differed significantly from foreign service officers. The term a_0 is the estimated intercept term for *all* respondents: the "shift" factors (technically, "dummy variables") have the value of 1 when a respondent is a member of the comparison group associated with the dummy variable and the value of 0 otherwise. If the intercept term for National War College respondents differs significantly from a_0, for example, this will show up as a significant value $p(t)$ for b_2; if so, the overall intercept will be a_0 (for foreign service officers) and respondents from the Office of Management and Budget) and $a_0 + b_2$ for the respondents from the National War College. If b_3 is also significant, then the overall intercept (applicable now only for foreign service officers) will be a_0, the intercept term for respondents from the National War College will be $a_0 + b_2$, and the intercept term for respondents from the Office of Management and Budget will now be $a_0 + b_3$.

The general equation is read, similarly, for the slope of the independent variable: b_1 is the overall slope estimate for all respondents, b_4 is the shift in the overall slope estimate for NWC, b_5 is the shift in the overall slope for OMB (i.e., the NWC slope is $b_1 + b_4$; the OMB slope is $b_1 + b_5$).

In analyzing the data all of these terms were entered into the computer as *potential* terms in a best-fit model. Stepwise regression was employed to select from these terms those which made a significant contribution in increasing R^2.[15] In other words, complexity was introduced conservatively up to the point where there was no significant increase in the proportion of the variance explained by a more complex equation.

The model of the mind used in this study has been a simple one. I have imagined that there is one "generic" mind, with each individual's mind

being a snapshot of a possible configuration of this mind. The simple equations in chapters 4 and 5 have implicitly assumed that all of these snapshots can be treated as equivalent except for the two independent variables under consideration. At the end of chapter 5 this simplifying assumption is removed and the best-fit linear additive model using all of the relevant variables from the study is presented.

One problem for further research is how empathy and explanatory power could be expanded by obtaining more data from more individuals. I am deeply aware of the simplicity of generalizing about all individuals taken together and I recognize that I have not differentiated among individuals in this study. There may be some individuals prone to use self-expression and projective intuitionism a great deal, others to use it little. It would be especially useful to study this phenomenon and determine what sorts of life experiences and professional training might contribute to unusual clarity, flexibility, and objectivity in the use of the imagination.

In addition, it is quite plausible that the minds of different people function in different ways; the opportunity to develop and test more complex models could increase explanatory power and insight. I have in mind the male sexual and narcissistic fantasies and fears in the foreign policy thought of ambitious men which, if my reading of the clinical literature of male narcissistic striving is correct, could clarify further some especially aggression- and fear-prone dynamics in international politics.

Validity of Results—General Considerations

The most serious question about the State Department study is, "Did these men tell the truth or were they trying to present positive faces of themselves to me and to the potential audience of the study?"[16] I am less concerned about the issue in the case of the military officers (who struck me as quite bold, frank, and forthright); and OMB specialists, who were chiefly concerned with economic problems and program evaluation in domestic affairs, would have little reason not to be candid. But diplomats are skilled at the gracious manipulation of their self-presentation and might be centrally concerned with putting forward a positive face.

In general I am inclined to believe that those who responded did so

candidly. The questionnaire was completed anonymously, did not deal with unusually sensitive issues, and the research was independent, done in association with a respected university, and had no official connection with bureaucratic superiors. Those who might feel pressured to alter responses had the far easier course of simply dropping the questionnaire into a wastebasket. Moreover each man could request, through a separate form, a report on the results of the study—the number returning the form was approximately equal to the number returning the questionnaire. Thus there was a minor incentive to be candid (i.e., the production of an accurate report), and I suspect fewer would have requested such a report if they felt they were providing biased data.

The strongest test for validity is whether the data obtained are internally consistent and converge with other data from other sources. The plausible interscale correlations reported in chapter 3 suggest either honesty or enormous and laborious systematic falsification; the first alternative seems far more likely. Compared with my expectations and prior knowledge I found no surprises, although there probably should be some adjustments, as discussed earlier in this appendix. The general portrait—well-educated, high self-esteem men, a basic assumption of the personal trustworthiness of others, at least moderately high ambition, a tendency toward competition, relative freedom from obvious neurotic symptoms—was what one would plausibly expect to find in the professional government elite of an advanced industrial country where both entrance and promotion are on the basis of merit. The picture may be slanted a bit toward the positive side but not, I think, excessively.

The only slight surprises encountered were on the policy side. Extrapolating from myself (as I conclude these men do as well), I was surprised to find as high a degree of tendency to use military force and to be fearful of Soviet ambitions as were reported by diplomats. I would also have expected efforts to present a "positive" face in the waning days of American involvement in Vietnam to have produced a low percentage of people who said they favored that war. But such a self-portrayal did not appear (almost two-thirds of diplomatic respondents and 83% of military officers said they would have favored an American war in Vietnam if they believed it would have succeeded by the end of 1968). Since my prior expectations

were not based on hard data, I am inclined to accept as reasonably valid the responses obtained here, with the reservations, expressed above, that some *increase* in the tendency to use force, in international mistrust, and in the fear of the Soviet Union should be made if the reader wishes to extrapolate from the sample to the State Department as a whole.

Good evidence for the validity of responses is the coherent pattern of correlations in the data that correspond with theoretical expectations. And these results, which show self-expressive generalization, mesh closely with the similar results obtained independently with the historical data of actual decisions between 1898 and 1968.[17]

One further question is whether the attitude responses given on the paper-and-pencil questionnaire employed here will predict to actual policy preferences in real world situations.[18] This general question can be decomposed into two subsidiary questions. First, will any of the attitude scales predict simply and directly to behavior? The answer is "probably not." In my view a field-theory orientation, a view of behavior as the result of *multiple* forces, considerations, and attitudes, is most appropriate for thinking about foreign policy behavior. A single predisposition is one, but only one, significant component. Second, do the attitude scales used efficiently represent such a significant component of the determinants of behavior in actual situations? This question cannot in the end be answered by paper-and-pencil tests, and this was one reason for the coordinate historical study, which found that dominance over subordinates, a trait associated with ambition, did indeed predict the use of force. But there are plausible reasons to think the answers reported do have substantial validity, especially for diplomats and probably for the military officers. These were well-educated and sophisticated professionals with considerable practice and experience (typically 15-20 years of career service) in facing the kinds of issues that were presented here, men working in an environment in which such issues are discussed both professionally and socially. Moreover the questions posed in the questionnaire were cast in terms of concrete policy situations and specific issues, close to real issues, and presented with "hard phrasing" rather than general beliefs or sloganistic sentiments (e.g., "A nation faced with a bunch of really dangerous enemies might be better off to shoot first and ask questions afterward.")[19] The attitudes of these men

were assessed by asking them fairly directly about their actual views developed and manifested in real-world situations.

The usual worry is that paper-and-pencil tests will be taken to predict more strongly to future behavior than is warranted. But, especially given the high percentage of cases explained by the personality trait of dominance over subordinates in the historical study, an opposite (and perhaps more relevant) validity question arises; namely, whether paper-and-pencil correlations obtained during a relatively tranquil period in international relations may *understate* seriously the effect of personality engagement on policy decisions in real-world decisions. It may be that emotional arousal and actually making decisions personally (often in a crisis atmosphere) decrease higher cognitive functioning and engage personal emotional dynamics with *greater* force than is captured fully by asking men sitting calmly at their desks, somewhat psychologically removed from these situations, what they would do.[20]

Validity of Results—Correction for Attenuation and the Conservatism of Results

The results reported in chapters 4 and 5 probably underestimate the impact of variations in personality on variations in policy and perception. Although a rigorous mathematical treatment is complex, a brief prose sketch of the problems of measurement error and of "attenuation" can outline the source of the underestimation.[21] It is generally known that a man's answer to questions tends to vary somewhat over even short periods of time (at least for the kinds of questions used in personality and attitude scales). The traditional theory of psychological measurement assumes that, if one is measuring an enduring personality trait or a meaningful attitude, then each individual can be thought to have one "true" score on the scale which is employed. The variability of the "observed" score from one testing to the next is conceptualized as "measurement error." (The mean of a man's observed scores across repeated tests is his true score; error terms are assumed to be normally distributed about the mean and to be uncorrelated with the true score.)

The reader will have noted that many of the relationships which have

been reported have a low R^2. In other words, variation in any one personality factor accounts, by itself, for only a small proportion of the variation in policy views. And, while I expect this is correct, the R^2 values which are reported could be biased downward substantially by the presence of measurement error of the type described. If we assume that the scales used in the study have a test-retest correlation of .75 (a reasonable assumption for scales of this type and length), then (in a bivariate regression) the R^2 value should be adjusted upwards by 7/9. The upward adjustment of R^2 tells us the proportion of variance in the dependent variable which we have explained out of the total variance we *could* explain. (The adjustment has the effect of subtracting from the observed variation an estimate of the amount of that variation due simply to measurement error).

As just illustrated, regression results are quite sensitive to measurement error. Not only is the computed R^2 biased toward zero, but the b coefficient also tends toward zero (and, in the hypothetical example above it should be increased by one third in absolute value), other estimates of error become too large, the computed t and F values are too low, and so forth.

I have not applied these corrections for attenuation to my results. In part this is because the rationales for such corrections are generally unfamiliar and employing them might make my argument appear self-serving and suspicious. In part it is because the n-dimensional mathematics involved in adjustment is complex and would not alter the conclusions I draw. The crucial point, however, is that measurement difficulties of the type I have described do not affect results in unknown ways. They *always* tend, in the bivariate case, to make the reported results an *underestimate* of the significance and strength of impact of the variables being tested, and thus the case for the conclusions is stronger than is apparent on the surface.