THE SCIENCE AND ART OF SURVEY CONSTRUCTION

Scales, Measurements and Human Responses

Dr. Bert Huggins
US Army Cadet Command

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JAN 2010		**00-00-2010 to 00-00-2010**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
The Science and Art of Survey Construction: Scales, Measurements and Human Responses	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **US Army Cadet Command,55 Patch Road,Fort Monroe,VA,23651**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Personnel and National Security: A Quantitative Approach (Unclass), 25-28 January 2010, Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**15**	

SCALES AND COMMON ERRORS

- Three major factors affect the validity of response scales:
  - Coding (scale agreement, expected sampling)
  - Human factors, aka, person-form effect (Cattell)
  - Clarity
- The most common form of scale used has been the Likert scale, and is the most susceptible to erroneous interpretation.
- The scale used determines the analysis that can be used, and the sampling size required to avoid Type II Error.
- The researcher should consider whether the response will be an independent or dependent variable in the analysis (or both).
- The researcher should consider normal rating (scales of 0-10 are logical and commonly considered, while a scale of 1-7 may infuse higher levels of response than expected).
- The age, and to a lesser degree the culture, of the intended audience should be considered.

SCALES COMMONLY USED IN SURVEYS

1. Nominal scale items: examples, race/ethnicity, gender, membership, region, etc.
2. Dichotomous scale items: yes/no, agree/disagree, etc.
3. Dichotomous scale or nominal scale with layers.
4. Ordinal scales, including most Likerts, order of merit.
5. Anchored Scales with numeric underpinning, with extreme values: Never/Always or Disagree Completely/Agree Completely.
6. Anchored scales spatially arranged along a single axis.
7. Spatial arrays with multiple axis.

Of the above, only 5-7 allow for parametric statistical manipulation.

SCALES AND COMMON ERRORS - CODING

Coding

- A common error in coding are to transfer ordinal data to sequential integer values.
- This is especially problematic with Likert scales, but can affect other response scales as well.
- A common scenario:

A Likert Scale has values of Dislike Very Much, Dislike Some, Neither Like Nor Dislike, Like Some, and Like Very Much. These are the responses, not a number associated by the researcher in coding. One can argue that the responses are ordinal, but the responses are frequently coded in the data as 1, 2, 3, 4, 5. Faulty assumptions are nearly impossible to avoid. Is the value of Dislike Some half of the value of Like Some? Can Like Some be equated to $4/5^{th}$ the value of Like Very Much? More to the point, for the individual respondent, is Like Some closer to Neither Like Nor Dislike or to Like Very Much. The respondent may respond Neither Like Nor Dislike because the question does not apply to them or they really do not understand the question. This should remove the individual to a separate category analyzed separately. That means that the scale is both ordinal and nominal.

In the end, coding these scales limits the analysis to the point that it is nearly ineffective. If the researcher understands this and does not intend to replace essentially categorical information with numbers, then there is no problem. Bar charts and simple descriptive statistics can still be used. CATREG can extrapolate some effects, but the sample sizes required to avoid Type II errors can be outside the limits of the research or higher than the population under study. However, this epiphany is altogether too rare.

LIKERT SCALE

36. Given blah, blah, blah, how do you feel?

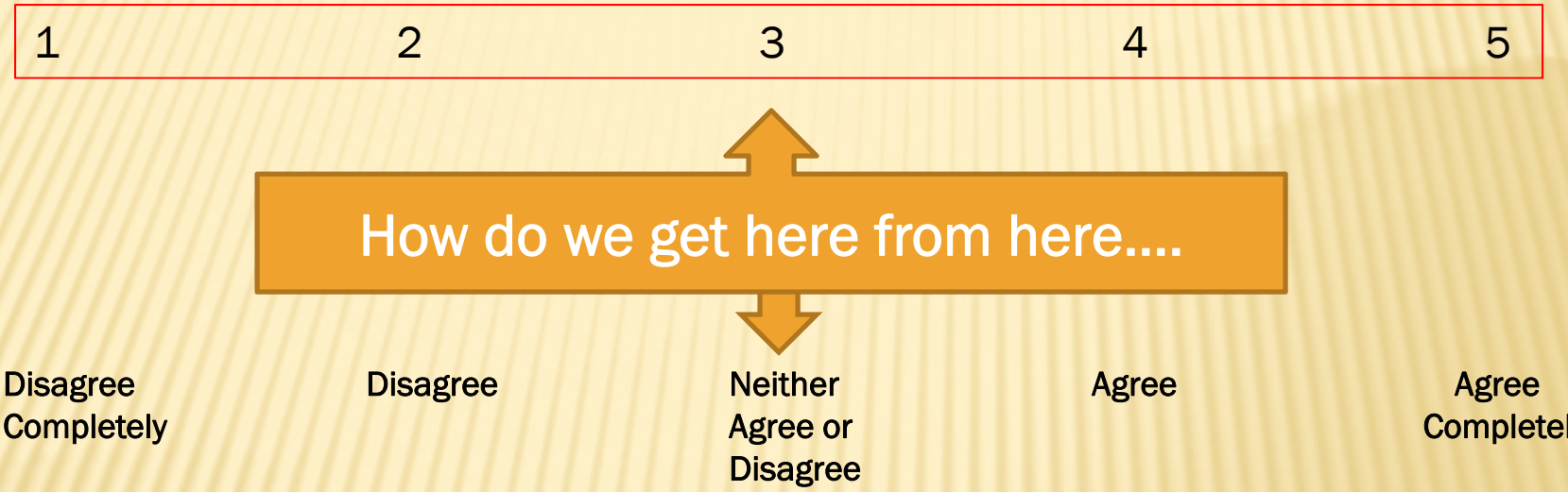
Disagree Completely	Disagree	Neither Agree or Disagree	Agree	Agree Completely
☐	☐	☒	☐	☐

What is this Guy trying to tell Us?

1) I got your feel right here.
2) I am incredibly bored with this.
3) I have no idea what you are asking.
4) If I check all middles, I can be done in no time.
5) Man, I am hungry.
6) I think I left the coffee pot on.
7) After careful consideration, I am right in the middle on this question.

- Before coding, some Likert scales present issues on response validity.
- Neutral points are not necessarily neutral.
- Our respondent on the left could be an outlier that should be treated separately, unless we are sure that his response reflects the answer 7).

CONVERTING WORDS TO NUMBERS



Scales associated with Likert responses are frequently expressed one through five, etc., and provide an illusion of numeric clarity.

The number scale above is elegant, simple and, unfortunately, fictitious.

Since its roots are words, which suffer from connotative dissonance (words are subject to interpretation, and interpretation leads to differences in responses, or error).

Even if we could assume, and we cannot, that everyone perceives the distance between disagree completely and disagree to be the same as between disagree and neither agree or disagree, the values coded cannot purely represent the response.

Some of the error can be mitigated by using zero as the mid-point, but this still assumes that the rater or respondent views equidistantly the points along the scale.

SCALES AND COMMON ERRORS - 2

Coding 2

- Sample sizes and the anticipated modeling/hypotheses testing.
- Use of binomial (dichotomous) variables as predictor variables.

The scales and resultant coding should be part of the plan of the end-state models and required hypothesis test. If the sample size is expected to be less than 1,000, the number of contributing factors that can establish the relationship between predictor data and dependent variable are extremely limited.

The rule of thumb: For any interval/ratio predictor variable in a linear regression formula, n > 30. For any dichotomous predictor variable N > 300. If the dichotomous variable is the dependent variable, n can be lowered, assuming relatively similar populations for 1 and 0. Logistic Regression is the formula of choice when Y is dichotomous. Error can occur, however, if Y is unevenly distributed between the two outcomes. The greater the inequity, the higher the chance of error. The only way to overcome the error is by oversampling. Therefore, the researcher should play out the possible models and hypotheses prior to determining what scales and what sample size may be required.

- A common scenario:

The research is about what factors contribute to bringing the prospect to sign a contract. The dichotomous dependent variable is the contract signature. A series of questions about the prospects' experiences that are expected to affect the decision are put into a survey. The prospects are asked to fill out the survey at some point prior to contracting. The survey has anchored scales with interval data, and some demographic information. The researcher will use none of the demographic data until after the initial analysis. However, only a fraction of the recruiters are engaged to do the survey. Most wait until very near the contract signing to have the prospect complete the survey. N = 830, but only 122 fail to contract. That effectively reduces the independent variables able to be simultaneously run to a handful. If the researcher wants to use demography, that segments n further. Lets say that the question comes up about Hispanic as a separate issue. Only 108 of the participants were Hispanic, and 78 of those contracted. Now the effective number of independent variables that can be safely considered in a regression is 1.

Dichotomous variables used in prediction sets are only plausible when there is approximately normal distribution expected in Y, or , the n is extremely high. Sample size must relational to the anticipated models, hypotheses, scales and anticipated analysis.

EXCEPTIONS TO DICHOTOMOUS VARIABLE UTILITY RULE

The more you know about the position of a particle, the less you know about its speed and direction, and vice versa.
Hiesenberg $\Delta X \Delta P \geq \frac{\hbar}{2}$.

- Pathways can be applied to surveys. In general, these are yes, no responses, which in turn lead to multiple options.
- The frequency of these options along with the depth of possible responses can lead to very small groupings, but very specific information on the individual. For example, if there are six layers of "yes" or "no" responses, you will create a potential for 2^6 groups (64). In order for this set of variables to be useful predictors in a normal distribution, the n $\geq$ 19,200; however, without equal distribution, it could be considerably larger.
- That said, the sophistication of pathways will lead to much more individual information.
- **In general, the more we know about one individual, the less we know about the group.**
- Surveys of this type are most appropriate for detailed information about a small sample.

SCALES AND COMMON ERRORS - 3

Human Factors

- Perceptions and interpretations variance between individuals
- General perceptual impact affecting responses

Scale interpretation is a common problem in responses. The most clearly worded question can be obfuscated by the scale associated. Rule of thumb: The more words, the greater the possibility of response error. This is one of the issues with Likert Scales that require the respondent to read and interpret the relatively weight of adjectives and adverbs (very, more, etc.) in combination with the term of measurement.

Rater Error: A greater issue is the differences between respondents based on age or culture. Youthful respondents (17-20) tend to respond in extremes (error of extreme measurements). Older respondents (24 and up) tend to respond more toward the middle (error of central tendency). Comparing these groups can lead to misunderstandings. The other problem with youth surveys is that the results will tend toward bi-modal distribution. That creates significant issues in interpretation of results with statistics that assume normal distribution. A Kolomogrov-Smirnov will identify anomalous distribution. Most parametric measurements suffer from lack of normal distribution. To enhance the opportunity for discerning patterns and contributions, expanding the scale will help. A five-point scale will generally result in more extremes (either 1 or 5), while a 10-11 point scale will have distributions of 1, 2, 3, or 6, 7, 8, 9. as well as 10).

General perceptual issues. Humans have an innate spatial recognition. Use of spatial scales can enhance response validity. Language and culture do not impact spatial. If words are used in the response scale, having the scale look more like an array than separate blocks will make use of this human trait. If blocks are used, numbers are better than words. Numeric misinterpretation is not as likely as verbal, but not quite as good as spatial. To a respondent, 4 is twice 2, and 8 is 4/5ths of 10. Test-retest reliability increases with numeric anchored scales, but spatial allows for an unlimited variance in response, and test-re-test exceeds all other scale instrumentation.

INTERPRETATION

What is this?

There is a hierarchy of clarity in human perception (visual). As we go down the hierarchical list, the agreement between any two individuals will dissipate.

That hierarchy includes:

1. Spatial – humans have innate spatial recognition that precedes language and abstract thought.
2. Figurative – recognition of representational symbols, pre-language.
3. Numeric – representation of collections, sizes or scales, pre-language.
4. Representational language, written.
5. Non-representational language.

Each level down represents a higher level of discernment and interpretation. Variance expands with each level. The degree to which questions and answers on a survey are interpretive, cultural, maturational and individual differences will overwhelm the clarity of response.

SCALES AND COMMON ERRORS - 4

Clarity

- Verbal clarity consistent with language and connotation
- Interactions of prior questions change the context of the response
- Interactions between response sets.

Clear and simple generally produces the higher resolution. Rules of thumb: 1) Almost never combine questions. If and is in the question or response, there may be problems. There are exceptions to this rule. 2) Avoid jargon, abbreviations or words potentially errant in connotation given the intended audience. 3) Parsimonious is always preferred.

A previous question can influence the response on subsequent questions, particularly with questions that have a personal or positive or negative stressing. One of the ways to counteract this is multiple questions within the domain addressing the same sub-element. Internal validity can be better assessed and chances of error reduced. The other way is to have multiple forms of the same instrument that change order; however, this can result in creating separate groups requiring separate analysis.

Jumping back and forth between scales has potential negative outcomes. Despite keeping the respondent on his toes, the chances of error increase by changing scales or direction of response.

DOMAINS AND BASICS OF CONSTRUCTION

The fundamentals of survey construction are more frequently ignored than followed.

- Build the survey based on domains, not items
- Build the survey based on what is needed to know everything to answer the underlying questions

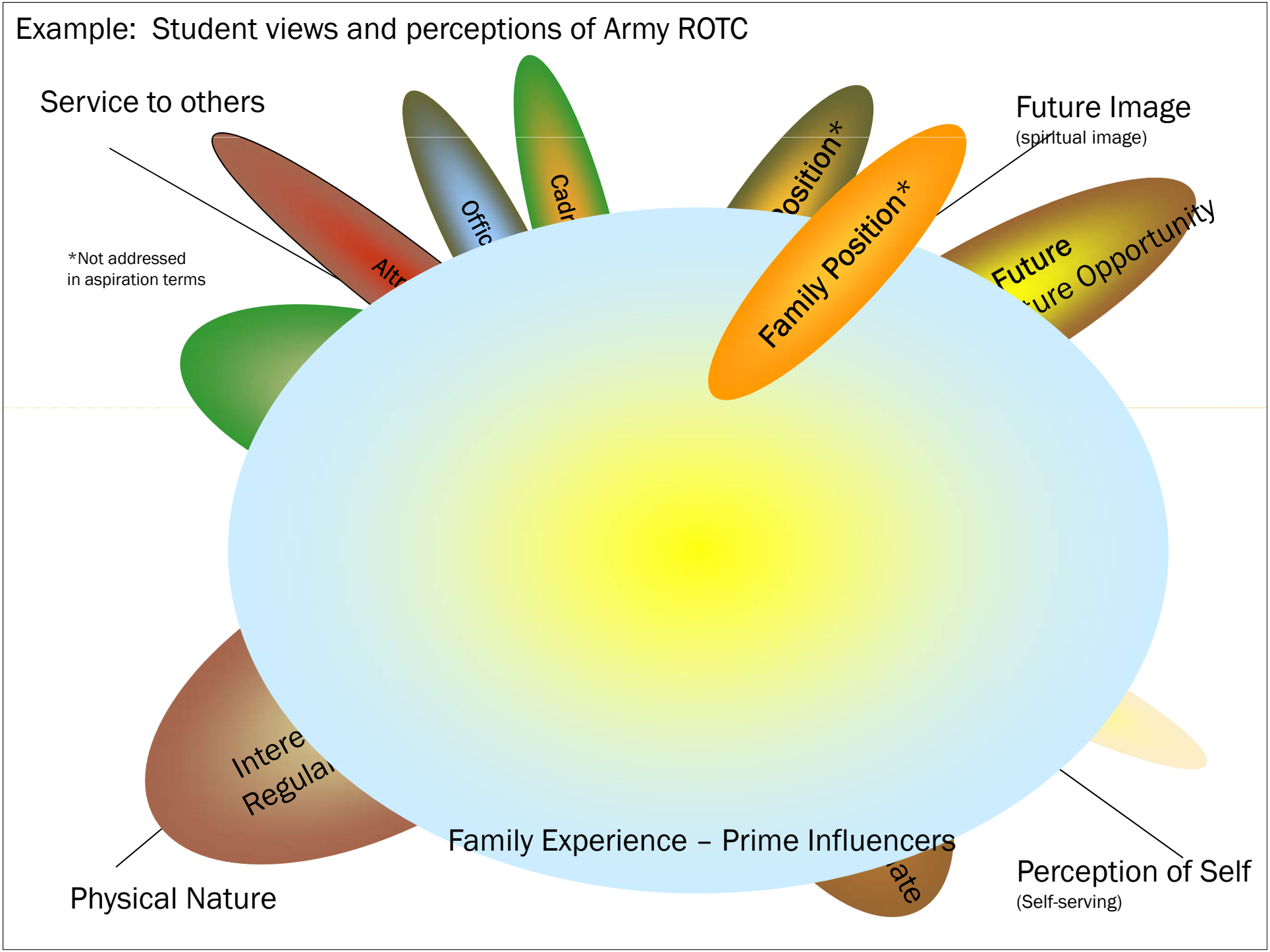
Survey domains should focus on the elemental aspects of human characteristics or constructs that can be expected, based on literature review, to contribute to differences in responses.

Essentially, anything that may make someone predisposed to answer the fundamental question should be part of the question process. For example, someone who is strongly adverse to exercise and physical exertion would not likely be interested in enlisting or participating in Army ROTC. However, if the superficial question of "are you interested in enlistment in the Army?" stands alone, the underlying reason(s) for either a positive or negative response would not be understood without asking the underlying question.

Multiple domains likely surround the simplest of survey questions. Multiple items need to be developed in order to clearly expose the nature of the domains. For all domains and items, a separate hypothesis needs to be developed.

In surveys: Fishing trips generally produce only an empty hook.

Example: Student views and perceptions of Army ROTC



BUILDING DOMAINS

Back engineer the domains from the hypotheses. If you don't have hypotheses, start over.

"You can't do what you don't know any more than you can come back from where you've never been." Joe Crosswell.

Domains should be seen as a set of known information and relationships, and theorized relationships. Questions, or items, should collectively describe the domain. Collectively, the domains should describe the viewpoint of the respondent anticipating the relationship to the dependent variable.

"If a hen and a half lays an egg and a half in a day and a half, how long does it take for a grasshopper to kick the seeds out of a dill pickle?"

Domains constitute the independent variables, and often proxy dependent variables. Without knowing what the independent and dependent variables or having a system of measurements, arriving at the correct response is unlikely.

RECOMMENDATIONS

Scales:

1) Optimal: Anchored scale with no numbers (requires measurements along an axis)
2) Preferred: Anchored scale (10 or 11 points minimum). The choice of anchors and whether the scale has negative and positive numbers is up to the type of question and intended interpretation.
3) Lesser value: Binomial or dichotomous variable (Yes, No//Applies, Does Not Apply)
4) Marginal value: 5 or 7 point Likert Scales, ordinal rankings, etc. (sometimes unavoidable, and sometimes useful as Y variables).
5) Nominal scales should be restricted to demographics, unless no other option is available, and only when sufficient sampling sizes permit non-parametric analysis or modeling.

The Good Idea Fairy is the enemy of all good surveys and scales.

Surveys do not say what someone really is, but what they perceive that they are.

The longer the survey, the more chance of non-completion, or random answers – shorten as much as possible without losing domain integrity.