V. MINIMIZING ERROR

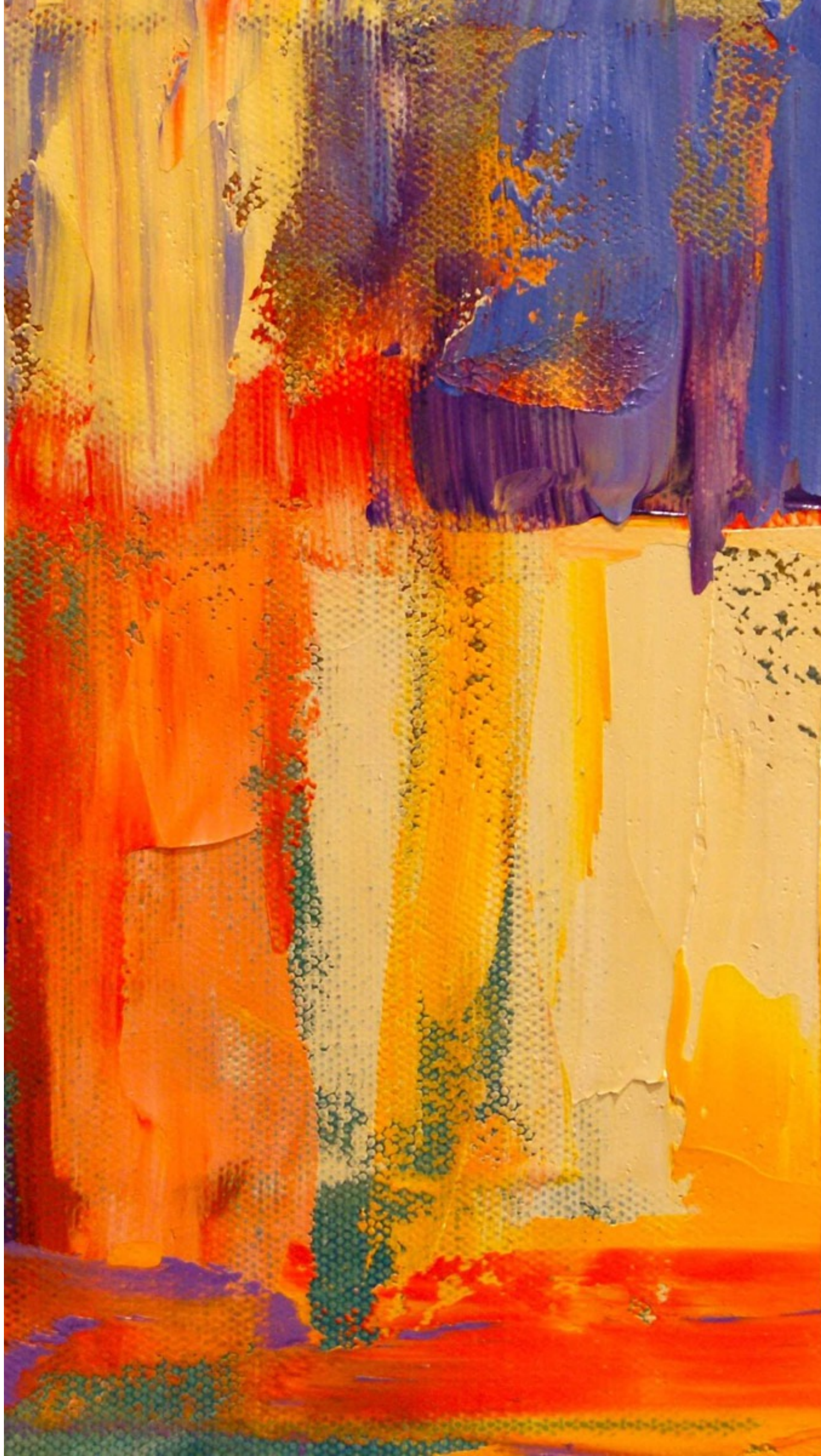
IN THIS CHAPTER:

- Possible sources of error due to:
  - Ethnographer bias
  - Informant bias
  - Coder bias
  - Poorly-designed coding rules
- Discussion of systematic error and random error
- How these two types of error may affect hypothesis tests
- Strategies for error minimization in (and during) research design

As has been discussed, a researcher must develop measures that represent the theoretical concepts she or he wants to investigate. However, even using measures with high validity cannot completely eliminate errors from the research process.

In this section we'll discuss other types of error, how they can affect your data, and how to best avoid them.

WHAT ARE THE ADDITIONAL SOURCES OF ERROR?



ETHNOGRAPHERS

Ethnographer bias may lead to the over-reporting or under-reporting of various cultural traits. For example, an ethnographer fascinated by religion might exaggerate the importance of ritual in his or her studied culture. Similarly, an ethnographer's preconception of a largely isolated culture may blind him or her to the essential role of trade in that society.

INFORMANTS

A person providing information to an ethnographer may give a somewhat **inaccurate report**. This could be purposeful, perhaps to minimize a behavior the informant(s) thinks will be viewed negatively, or it could be that the informant is not particularly knowledgeable about the domain of interest. Regardless, it will affect the quality of data.

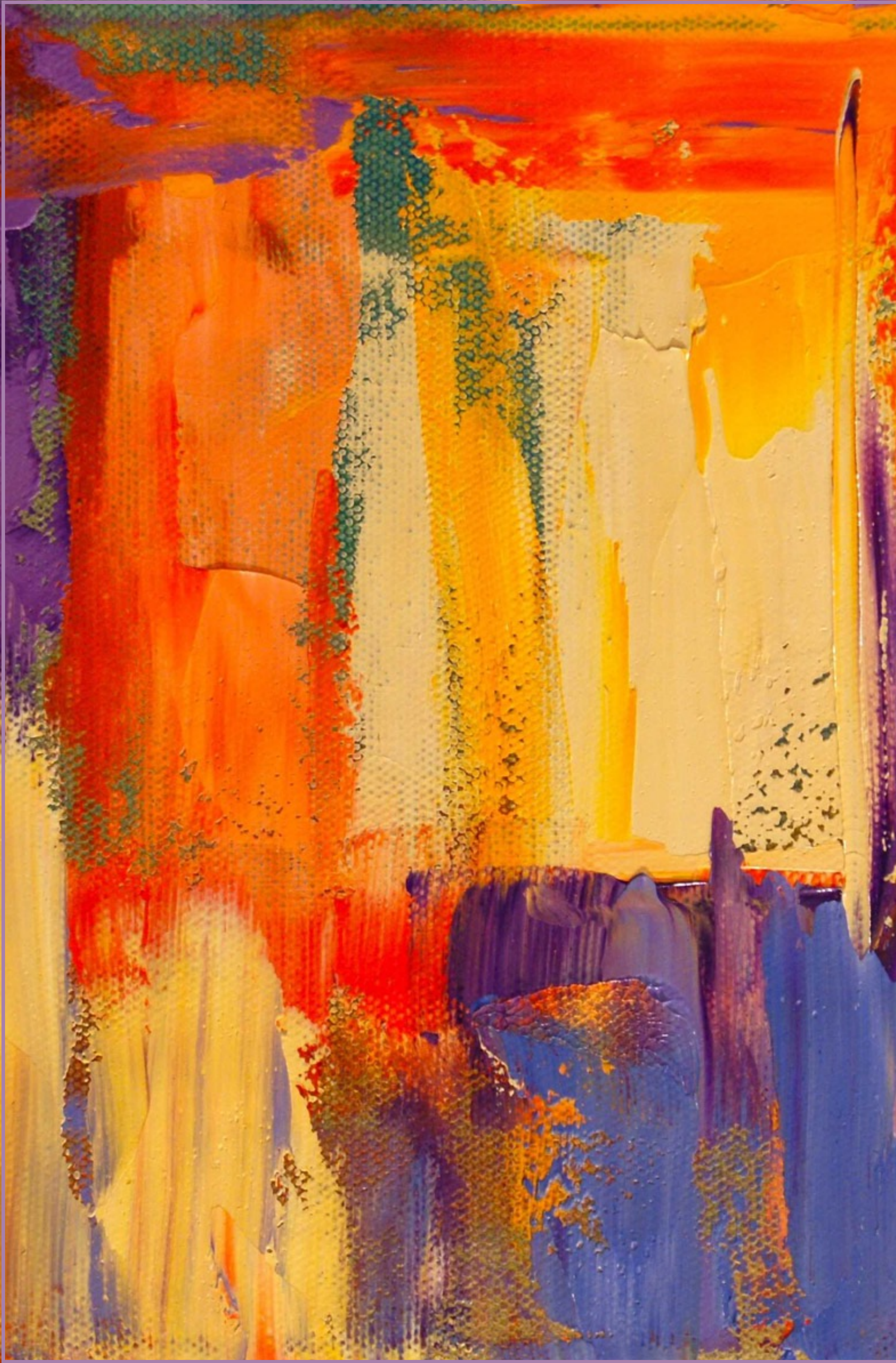
CODERS

Sometimes coders unknowingly rate societies according to their **biases**. For instance, data will be missed if coders focus on the cultural practices that seem most familiar to them.

CODING RULES

It will be difficult to accurately represent data with **poorly-designed codes**. Coding rules can be problematic in various ways: they might **exaggerate** minor differences between cultures, **obscure** major differences, or attempt to represent too many independent cultural traits in one measure.

TYPES OF ERROR
&
HOW RESULTS MAY BE AFFECTED

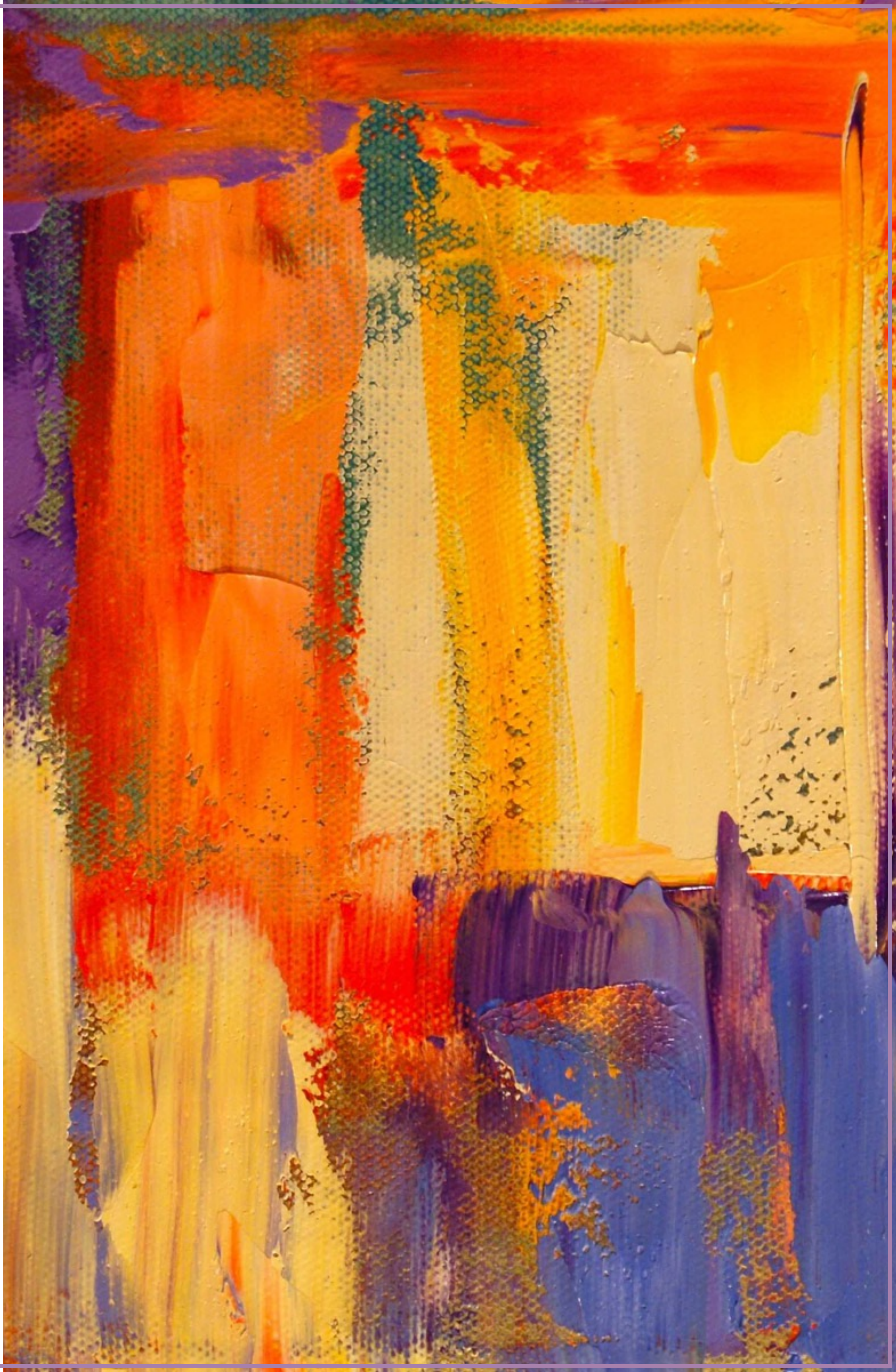


SYSTEMATIC ERROR

Systematic error means that there is a bias in the way data is **reported, measured, or coded**. In other words, all cases that deviate from the true score are affected by the same error. One or more of your measures is off by a constant. Therefore, affected cases in a data set will either all be **inflated** or all be **deflated**.

RANDOM ERROR

One or more of your measures is off but **not by a constant**. An affected case in the data set could be either inflated or deflated. Because **random error is unbiased**, any two cases could deviate from the true score in different directions and amounts.



The consequences of error will depend on the type of error and how many variables are affected. Let's say you have variables A and B, and you're investigating their relationship:

- If **ONLY variable A** is systematically altered, the association between variables A and B will not change; the two variables still predict each other to the same degree.
- If **both variable A and variable B** are systematically altered (known as **double systematic error**), the consequences depend on whether the systematic errors on variables A and B are correlated. If they are, double systematic error could produce a modest false relationship. If the errors are not correlated, a false relationship is unlikely to be created. If the errors are not correlated and consistent, a true relationship is unlikely to be obscured.
- If **random error occurs in one or more variables,** the errors may cancel each other out; the average value should be largely unaffected. If the errors affect the association, random error will almost certainly make a true relationship look weaker.

GENERALLY, BOTH SYSTEMATIC AND RANDOM ERROR DECREASE THE LIKELIHOOD OF FINDING SUPPORT FOR AN ASSOCIATION. IN SHORT, IT IS MUCH MORE LIKELY FOR ERROR TO OBSCURE A VALID FINDING THAN YIELD A FALSE RESULT.

HOW TO MINIMIZE ERROR

WHEN PICKING VARIABLES:

- **Maximize validity**.

  This means **maximizing the fit between the theoretical concept and the measure**. Different types of validity were discussed in the previous section; feel free to review them if necessary.

- **Minimize the amount of information that a coder must infer**.

  In coding, variables fall along a spectrum from low inference to high inference. **Low inference variables** tend to deal with cultural traits that are conceptually straightforward and clearly observable (e.g. the shape of dwellings); **high inference variables** require a more complicated coding decision on the coder's part. When you choose variables to code, high inference variables such as "evaluation of women" can be broken down into lower inference variables such as "presence of bride price." Lower levels of inference will increase the accuracy of coding decisions and agreement between coders.

WHEN SELECTING DATA:

Decide your time and place focuses. For synchronic cross-cultural comparisons, all variables should be measured at the same time period for each case. Consider these ways of picking a time/place focus for data:

- On theoretical grounds: for example, wishing to test a variable for time periods before European contact
- On measurement grounds: for example, one ethnographer provides substantially more detailed information
- Based on sampling criteria: if using data from a particular sample, you will need to match the time/place focus of that sample

If two or more variables cannot be measured from the same time period, you might:

- Decide on an acceptable time range (e.g. +/- 10 or 15 years surrounding the ethnographic date) and exclude cases outside that time range.
- Code all cases but include a variable that identifies how well the data matches the time focus. This will allow you to run analyses both with and without the cases outside the time range.

WHEN DESIGNING CODES:

- **Make a code sheet with clear explanation of variables, codes, and steps** needed to make coding decisions. Specifics about the code sheet can be found in Chapter 4.
- **Give instructions about inferring the absence of a trait.** The coder should only infer the absence of a trait when there is sufficient information on the broader topic (e.g. discussion of marriage arrangements with no mention of bride price).
- **Include data quality codes.** If data quality is assessed, coders can rate the quantity and quality of information provided by the ethnographer. Is information stated clearly and unambiguously? Is data generalizable to the whole population, or is it an anecdotal account specific to one informant or family? Once all of your cases are coded, you may choose to run tests without the cases with low-quality data. Exclusion of such cases might yield stronger results.
- **Evaluate your coding scheme by pretesting cases that are not in your study.** Have other people attempt to code those cases and discuss any discrepancies. This step is essential to ensure that you can get quality data from the coding process. Pretesting often highlights weaknesses in a coding scheme and provides opportunities to revise and improve your codes.

AFTER CODING HAS BEGUN:

INTRODUCE DATA-QUALITY CONTROLS.

This strategy is one way to counter potential ethnographer and/or informant error.

It requires additional time and resources, so we consider it optional. Naroll (1962, 1970) suggests that when researchers run statistical tests on coded variables, they control for data-quality variables such as:

- The length of time an ethnographer spent in the field
- Whether the ethnographer spoke the local language
- The ethnographer's gender

The idea is that one or more data-quality variables could be driving your substantive findings. However, Ember et al. (1991) only suggest the use of this method when researchers have reason to suspect that a data quality control is related to a main variable. Including data quality variables in your coding scheme (discussed above) directly related to each variable is a more convenient and potentially more relevant method for checking data quality.

MAXIMIZE RELIABILITY

Reliability can be thought of as **consistency or stability** in measurement.

In other words, we want different coders to have a **high degree of agreement** in their coding.

Please see the next section for more details on reliability.

SUMMARY

- **Maximizing validity** is one of the best ways to minimize error. However, there will always be possible error from ethnographers, informants, and coders.
- Ways to minimize these types of errors include:
    - Strive for **low inference variables** that **maximize inter-coder reliability**
    - Decide on a **clear time and place focus** and give coders appropriate sources for that focus
    - Give clear **instructions about when to allow inferred absence** of a trait
    - Include a **data quality score** for every coded variable so that lower quality scores can be eliminated in later analyses if necessary
    - **Pretest your coding scheme** with multiple cases and coders
    - Consider any other **data quality controls**, such as degree of ethnographer familiarity with the native language

REFERENCES

Ember, Carol R., Marc H. Ross, Michael Burton, and Candice Bradley. 1991. Problems of Measurement in Cross-Cultural Research Using Secondary Data. Special Issue, Cross-Cultural and Comparative Research: Theory and Method. *Cross-Cultural Research* 25:187-216.

CITATION AND TERMS OF USE

Using this course: The material in this course is intended for educational purposes only--either by individuals for personal use or for classroom use by instructors with appropriate attribution. For any commercial use or other uses please contact HRAF. (Email: <hraf@yale.edu>).

This chapter is from: Carol R. Ember. 2016. Introducing Cross-Cultural Research. Human Relations Area Files. <http://hraf@yale.edu/ccc/>