Science Education Collection

Reliability in Psychology Experiments

URL: http://www.jove.com/science-education/10046

Overview

Source: Laboratories of Gary Lewandowski, Dave Strohmetz, and Natalie Ciarocco—Monmouth University

In order to study something scientifically, a researcher needs to determine a way to quantify it. However, psychological constructs can be challenging to measure and quantify. This video examines reliability in the context of content analysis.

A recent study in the journal Pediatrics reported that 4-year-olds who watched a fast-paced cartoon had worse performance on cognitive tasks, such as following rules in a game, listening to direction from an adult, and delaying gratification, compared to other children who watched a slower paced cartoon.[1] In addition to the pace of the cartoon, the content of the cartoon may also have deleterious effects on its young viewers.

This video uses a simple two-group design, to exemplify the issue of reliability, in examining the question of whether the cartoon SpongeBob SquarePants has more inappropriate content than does the cartoon Caillou.

Procedure

1. Define key variables.

1. Create an operational definition (*i.e.*, a clear description of exactly what a researcher means by a concept) of inappropriate content.
2. Consult definitions created by the organization TV Parental Guidelines and approved by the Federal Communications Commission.
3. Inappropriate content is any crude or rude behavior (*e.g.*, toilet humor), depictions of verbal or physical aggression (*e.g.*, name calling, hitting, *etc.*), bad language (*e.g.*, curse words), or references to drug use, violence, or sex.

2. Create coding categories from the operational definition of inappropriate content.

1. Design a systematic process for the research participants (referred to here as the raters) to identify instances of the targeted, inappropriate behavior (see **Table 1**).

Coding Categories	Themes and Exemplars	Count
Crude Behavior	Toilet humor Purposefully disgusting behaviors	
Rude Behavior	Disrupting others Poor Manners	
Language	Using curse words	
Verbal Aggression	Insults Yelling Name-Calling	
Physical Aggression	Hitting Pushing/Shoving Tripping	
Drug References	Verbal (suggestive statements/conversation) Nonverbal (mimicking drug use)	
Sexual References	Verbal (suggestive statements/conversation) Nonverbal (mimicking sexual acts)	

Table 1. Example of how to record instances of inappropriate behaviors. This log can be systematically used across raters.

3. Instruct raters to separately watch the same episode of SpongeBob SquarePants and provide coding counts.

4. Instruct raters to separately watch the same episode of Caillou and provide coding counts.

5. Compare ratings to see if the Raters came up with similar ratings for each show.

1. Reliability is the ability to consistently measure the variable—inappropriate content.
2. Inter-rater reliability is the ability for more than one person to measure the variables and for their measurements to be in accord.

Results

The results indicate that the raters had a high level of agreement or consistency in their ratings within each cartoon episode, which indicates high inter-rater reliability (**Figure 1**). There is also reliability or consistency in SpongeBob SquarePants episodes having more inappropriate content than Caillou. The results also revealed individual biases amongst raters. For example, Rater 3 reported more inappropriate content in SpongeBob than the other 2 raters, and Rater 1 reported less in Caillou than other raters.

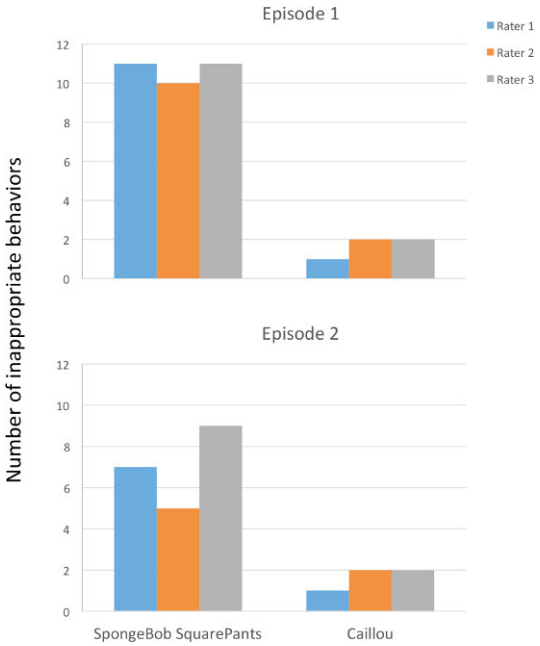

Figure 1. Instances of inappropriate content by rater and cartoon for episodes 1 (top) and 2 (bottom).

Applications and Summary

Researchers have increasingly turned their attention toward analyzing television's content, especially as it relates to children. As discussed prior to this current experiment, a recent study in the journal *Pediatrics* correlated the fast pace of the SpongeBob SquarePants cartoon to relatively poor cognitive abilities in the children who watch it.

Since the results of our experiment appear reliable, future research could examine whether the relative amount of inappropriate content in SpongeBob is also (or alternatively) responsible for children's lower cognitive performance after watching.

One of the most important applications of reliability is in the use of survey instruments. Researchers must be sure that participants will consistently answer each of the items in a particular scale. That is, in a 5-item measure of life satisfaction, participants should answer items 1 and 2 in a somewhat similar fashion to how they answer questions 3, 4, and 5. In addition, researchers want to make sure that their measurements in an experiment are consistent over time. So if a researcher is using pupil dilation to indicate interest in a stimulus, the researcher must be sure that pupil dilation is a consistent indicator of interest.

References

1. Lillard, A. S., & Peterson, J. The Immediate Impact of Different Types of Television on Young Children's Executive Function. *Pediatrics.* **128**(4):644-9. doi: *10.1542/peds.2010-1919* (2011).