Humanities and Social Science Research; Vol. 8, No. 1; 2025
ISSN 2576-3024 E-ISSN 2576-3032
https://doi.org/10.30560/hssr.v8n1p32

A Comparison of Methods for Binning Responses to Open-Ended Survey Items About Everyday Events: A New Tailored-Binning Approach

Derrick M. Angier[1] & John D. Mayer[1]

[1] Psychology Department, University of New Hampshire, United States

Correspondence: Derrick M. Angier, Psychology Department, University of New Hampshire, Durham 03824, United States. Tel: 1-802-989-5155. E-mail: derrick.angier@unh.edu

Received: January 10, 2025; Accepted: February 4, 2025; Published: February 9, 2025

The authors report no financial or other conflicts of interest in the writing of this article.

Abstract

Psychologists and other researchers employ diverse types of surveys and survey items. *Lifespace*-type items include questions about a person's external life conditions and activities, for example, "How many hours did you spend playing video games last week?" and elicit countable responses, for example, 0 hours, < 1 hour, 1-2 hours, 3-5 hours, etc., where the response intervals are referred to as bins. We report on a procedure for creating tailored bins for participants and compare it to two other binning approaches in an original and a replication study (*N*s = 263 and 246). The Tailored binning approach developed here is highly structured, but, for convenience, can also be applied in less formal fashions. In the structured form focused on here, it compares favorably with alternative approaches such as Equal Percentage and Equal Interval methods. The pros and cons of each method are described in the Discussion and recommendations are provided.

Keywords: binning, scale development, biographical data, lifespace

1. Introduction

1.1 Lifespace and Biographical Data

Psychologists often employ test-scale items that ask participants to describe their life status and activities, such as the number of conversations they had with friends on a particular day, the number of books they own, or the length of their commute to work. We refer to the broad class of such items here as lifespace items (Brackett & Mayer, 2006; Ivcevic & Mayer, 2009; Mayer et al., 1998; Perkey et al., 2018). This class of data is one of the key approaches to personality measurement of the 20th century (Mayer & Bryan, 2024) and is an umbrella classification that includes biographical data (biodata) (Mael, 1991; Sulastri et al., 2015), act-frequency data (Buss & Craik, 1983; Chapman & Goldberg, 2017; Ivcevic, 2007), and behavior data (Paunonen, 2003). Lifespace data can be contrasted with reports of self-judgments such as "I am a dependable person", in that lifespace items are, in principle, externally verifiable and draw on less subjective information (Buss & Craik, 1983; Ivcevic & Mayer, 2009; Mael, 1991; Soto & John, 2017). Lifespace items such as "How many years of college have you completed?" do not require interpersonal comparisons, only a recall of past and present experiences and behaviors (Mael & Hirsch, 1993).

1.2 Binning Item Responses

When creating a measure that employs lifespace items, estimating the range of possible responses ahead of their actual administration can be challenging (Borgers et al., 2004; Simms et al., 2019). For that reason, researchers may collect open-ended numerical responses to a survey in its early form. Although those numerical responses could be analyzed more-or-less as collected, doing so would run the risk of including potential outliers, including keystroke errors that can inflate or minimize behaviors, and highly unequal intervals. Furthermore, it can be useful to bin similar response values together into closed-ended alternatives, rather than maintain open-responding, for the clarity and simplicity it offers respondents.

Binning (sometimes referred to as discretization or bucketing) refers to the post hoc creation of categories from participants responses; for example, by dividing the responses into several intervals of equal size as originally proposed by Sturges (Sturges, 1926), or using more contemporary techniques (De Brouwer, 2021; Freedman & Diaconis, 1981; Prieto Curiel et al., 2020). An effective binning method can represent data accurately, in an organized fashion, and may be used as the basis for the creation of response options for future versions of an open-ended scale, for example, by creating dropdown menus with fixed-interval choice alternatives based on the original bins, thereby laying the groundwork for a subsequent survey revision. However, there are many different ways to bin and it may not be clear which method is most effective for a given purpose. We discuss several approaches to binning and introduce a tailored approach that may be of particular use in many circumstances.

1.3 Overview of Binning Methods

There are several common approaches to binning. An initial division might be into the (a) statistical/formulaic such as "Equal Percentage" in which each bin contains an equal percent of responses, and (b) tailored, which involves some researcher-judgment imposed on the binning method; Tailored binning sometimes is referred to as "supervised" in the machine learning context (Dash et al., 2011), but here we focus on statistical and tailored methods rather than machine-based learning.

Statistical/formulaic binning itself chiefly divides into Equal Percentage and Equal Interval techniques (Kotsiantis & Kanellopoulos, 2006; Lavielle & Bleakley, 2011); however, less-common unsupervised methods exist, such as cluster-based binning methods, which tend to require a precise criterion of *k* categories or clusters to function (Dash et al., 2011). The present work examines the two chief statistical/formulaic approaches to binning—discussed next—and compares them to a Tailored binning approach. The subsequent empirical example work examines a questionnaire about online gaming, and our examples will be drawn from that work (Angier, 2021; 2024).

Equal Percentage Binning

Equal Percentage binning (also referred to as equal-size or equal-frequency) divides participants' responses into each of *k* categories containing an equal percentage of responses. Say that 100 participants answer the question "How many times did you play a videogame last week?" with a range of answers from zero to 140. Creating bins using equal percentages would entail, for example, creating five bins in which 20% of the responses end up in each bin (or ten bins of 10%, etc.). One can specify either the percentage per bin, which dictates the number of bins, or specify the number of bins, which dictates the percentage (e.g., 10 bins yield 10%) (Dash et al., 2011; Lavielle & Bleakley, 2011). A commonly used approach to accomplishing this is to rank-order the responses and to place values into a bin until the bin reaches or exceeds the allotted percentage and then proceed to form the next bin. Note, all future references to Equal Percentage binning here will refer to this version.

The percentages in a given bin may be imperfectly divided owing to skewed distributions, which often arise in frequency-based responding; reducing skewness is relevant to the comparisons we draw among methods. For instance, in a case of five bins each targeting 20% of the responses, if the behavior inquired about is unusual or very specific in the sample, such as times played the video game "League of Legends", the first bin of zero might contain 50% or more responses while subsequent bins might require a large range, e.g., from "played 1 to 10 times" to meet the 20% criterion if each "time played" individually represented a small proportion of the responses.

Equal Interval Binning

Researchers who employ Equal Interval binning first calculate the total range of values reported by participants and then divide the range into *k* categories of equal intervals. One might begin with a general "play video games" question, but this time specify *k* = five categories. If the obtained responses range from zero to 25, for example, the equal intervals could be the five bins of: 0 to 5, 6 to 10, 11 to 15, 16 to 20, and 21 to 25 (Lavielle & Bleakley, 2011; Sturges, 1926). If we assume the skewed distribution of our "League of Legends" example then the vast majority of responses could fall into the first bin (0 to 5), while the other bins may be notably sparse. Approaches for optimizing the number or width of bins are available (Doane, 1976; Freedman & Diaconis, 1981; Scott, 1979; Sturges, 1926).

Tailored/Logical Modifications of Binning Categories

As useful as Equal Percentage and Equal Interval binning are in general, there are many special cases where research intervention may improve the bins utilized for a particular item. Researchers can, for example, choose meaningful starting places for the bins, or meaningful numbers of bins on a per-item basis. An example of starting at a non-zero interval is the US Census data for employed workers' salaries (Semega et al., 2020). The Census Bureau recognizes that child labor under the age of 15 is illegal and begins with the 15 to 24-years-of-age interval rather than an age of zero. As a second example, the government tax code regards salaries of less than $53,700—an approximately $55,000 range—as subject to little taxation, whereas those who earn more are depicted in broader intervals, for example, salaries from $207,351 to $518,400—a roughly $300,000 range— are grouped together, about 18 times the range of the lower salary level (Current Population Survey: 2019 ASEC Technical Documentation, 2020).

Tailored Binning

The Tailored binning process we introduce here is intended for application to general lifespace data. This broad approach is intended to satisfy several criteria for binning, described below.

1.4 Evaluative Criteria for Binning

Prior literature suggests several criteria by which to evaluate the performance of a given binning method. A specific approach to binning can be evaluated to the extent that it meets five criteria, including that the method:

(a) Creates between four and eight bins for the participant responses to a given item. A higher number than eight may create “nuisance bins” that contain no responses and add unnecessary cognitive load for survey-takers. Research with Likert scales indicates that increasing the number of response options beyond six provides limited to no benefit (Simms et al., 2019). Alternatively, too few bins can obscure important information if they combine categories that have meaningful distinctions between behaviors (Buss & Craik, 1983; Mael, 1991) and can annoy participants who wish to make finer distinctions; in studies of children and adolescents, responses with four categories appeared optimal (Borgers et al., 2004).

(b) Reduces the impact of outliers by merging them with bins with more common and plausible responses. Outliers may erroneously affect the overall distribution and results of item responses and/or reflect keyboarding errors (Myers et al., 2013), akin to Winsorization (Tukey, 1962).

(c) Allows for as many populated bins as possible: Empty (i.e., unused) bins not only create unnecessary cognitive load on participants but may interfere with data analysis as some statistical techniques perform less well as the number of empty bins increases (Prieto Curiel et al., 2020).

(d) Reduces skewness in the response distributions for items to reduce statistical issues in later analyses.

(e) Provides additional tailoring, where applicable, to retain the meaningfulness of responses; for example, if reports of zero, one, or two hours a day could be qualitatively large and informative, keeping the values of “0”, “1”, and “2” as individual and distinct bins rather than merging them.

(f) Provides a template for closed-ended response categories for use in subsequent revisions of the survey. Choosing a good binning method for later closed-ended responding assumes that the relative performance of the method is stable across samples. We will examine the degree to which the performance of a set of bins exhibits stability from Study 1 to Study 2 in the present research.

1.5 Overview of the Present Research

The present work draws on two samples of data pertaining to online gaming that included a lifespace survey that had employed open-ended responses (Angier, 2021; 2024). The purpose of the studies was to examine personality in relation to gaming behavior, but here the focus is entirely on comparative approaches to binning data from the lifespace measure. To illustrate and compare Equal Interval, Equal Percentage, and Tailored binning, we apply the three methods here to the two samples drawn from that work (N = 263 and 246). Although the lifespace questions employed here focused on gaming, they are in many ways representative of lifespace questions more generally as they ask for theoretically verifiable and objective information, such as asking for the hours spent engaged on a task, the number of times performing an action, and the number of objects in the environment.

For Study 1 (N = 263), each open-ended item was binned according to the rules of Equal Interval, Equal Percentage, and Tailored binning, and the outcomes compared. For Study 2 (N = 246), the bins established in Study 1 were applied without change to the still open-ended data from Study 2 to check their consistency of performance across samples.

2. Study 1

2.1 Study 1 Hypotheses

In Study 1, we compared Tailored binning to Equal Percentage and Equal Interval methods, to test the following:

Hypothesis 1. Tailored binning would yield item distributions closer to normality (less skewed) than the other methods. We tested this hypothesis by calculating the skewness for each item for each binning method and then comparing them statistically.

Hypothesis 2. Tailored binning would produce an increased number and proportion of populated (non-empty) bins per item relative to Equal Percentage and Equal Interval binning. Conversely, there will be fewer empty bins. We tested this by applying tests for differences in proportions (of populated and of empty bins) across methods.

Hypothesis 3. Tailored binning would produce a smaller proportion of items that contain empty bins relative to Equal Percentage and Equal Interval binning. (This is distinct from Hypothesis 2 in that it examines *items* with or without empty bins as opposed to the overall number of empty bins averaged across items).

To assess this hypothesis, we tested the differences in the proportion of items that contained an empty bin, across binning approaches.

3. Study 1 Method

3.1 Participants

The Study 1 sample consisted of 263 undergraduate students (100 male; 162 female; 1 other) with a mean age of 19.10 and were recruited through the UNH Psychology SONA study recruitment system. The University of New Hampshire is a public university in New England with students that are approximately 83.9% White/European American, 4.1% Asian/Asian-American, and 12.0% other races/ethnicities. Screening resulted in the removal of 42 original participants for such issues as overly quick responding or excess omitted data and is reported in detail elsewhere (Angier, 2021).

3.2 Methods and Measure

The To-Be-Binned Scale: A Gaming Lifespace Survey

A 52-item version of a lifespace survey that assesses video gamers' gaming-related behaviors and preferences was used in this binning study. Sample questions included "How many times in the last week did you go to bed immediately after playing video games?" and "How many times in the last week did you communicate a positive greeting to others at the beginning of a round/match/game?" All items called for open-ended numerical responses, making them ideal for a study of binning. The scale was included as part of a broader survey concerning gaming behavior and personality and is reported elsewhere (Angier, 2021).

All 263 respondents of this lifespace survey were used in the binning and analyses. The survey, however, is a branching survey and 99 participants answered just 41 of the 52 items. We report data for all the items, acknowledging that the *N* dropped from 263 to 164 for 11 items.

Binning Procedures

For the purposes here, participants' open-ended responses to the 52 items of the lifespace survey were binned in three ways: Equal Percentage, Equal Interval, and Tailored methods, with a target of between 4 to 8 categories for each method, and selecting the maximum categories (8) within that range where possible.

Equal Percentage

Equal Percentage binning was carried out using the SPSS Visual Binning tool, so-called because it creates a visual histogram in a window from which one can adjust settings to bin items (Myers et al., 2013). Bins were set at a maximum of eight by specifying 12.5% of the responses to go into each bin—or as close as possible. The procedure uses the minimum number of values needed to meet or exceed that specified percentage and then continues to the next bin until all values are binned. This can (and did) often leave fewer than eight total bins for responses to an item if a specific response occurred more than 12.5% of the time.

Equal Interval

Equal Interval binning was again carried out using SPSS, utilizing syntax to establish the interval size appropriate for eight bins for each item by taking the item's range and dividing by eight. The established SPSS algorithm always created eight bins as long as the range of response values was eight or more. The method then placed the responses as evenly as possible into those equal width bins. For those items for which participants entered fewer than eight responses, the resulting number of bins was adjusted to match the maximum number of different responses (e.g., 5 total bins for an item that yielded responses of just 0, 1, 2, 3, and 4).

3.3 Tailored Binning & Procedure

A visual guide to the Tailored binning method introduced here in Figure 1. In "Part A" (Figure 1, top), each item is evaluated according to the number of bins appropriate to it. If fewer than 4 responses were given, then the decision tree branches to the left. If 4-8 responses were provided, the specific number of bins was accepted as a starting point. If a large number of responses was given, then a larger number of bins was employed to start (Figure 1, right). A logical floor or ceiling value also was applied if applicable (e.g., adding a bin with "0" as a response option if that is a currently unselected possible response), and counting the number of bins that are now available for that item. If the number of bins is within the 4-to-8 range criterion then binning is complete.

If the bins are too few (Figure 1, Part B), then the researcher chooses whether to keep the number of bins below the criterion as-is or to increase the number of bins until a criterion of 4-8 bins is met. To increase the number of bins, empty bins are added between two chosen response values or at the extreme end of responses (Figure 1, Part B) depending on what makes most sense.

If the bins are too many, they can be combined—but before doing so, they are checked for skewness (Figure 1, Part S). If skewness is substantial for that item, then the researcher may proceed to make adjustments, for example, merging low response-rate bins at the skewed tail(s). If skewness is not a concern, and the number of bins is within the criterion, then binning is complete.

If there remain more than eight bins and the skewness adjustment of Part S was already completed or deemed unnecessary, then the steps of Part C were followed to merge the bins until the number of bins reaches criterion (Figure 1, Part C).

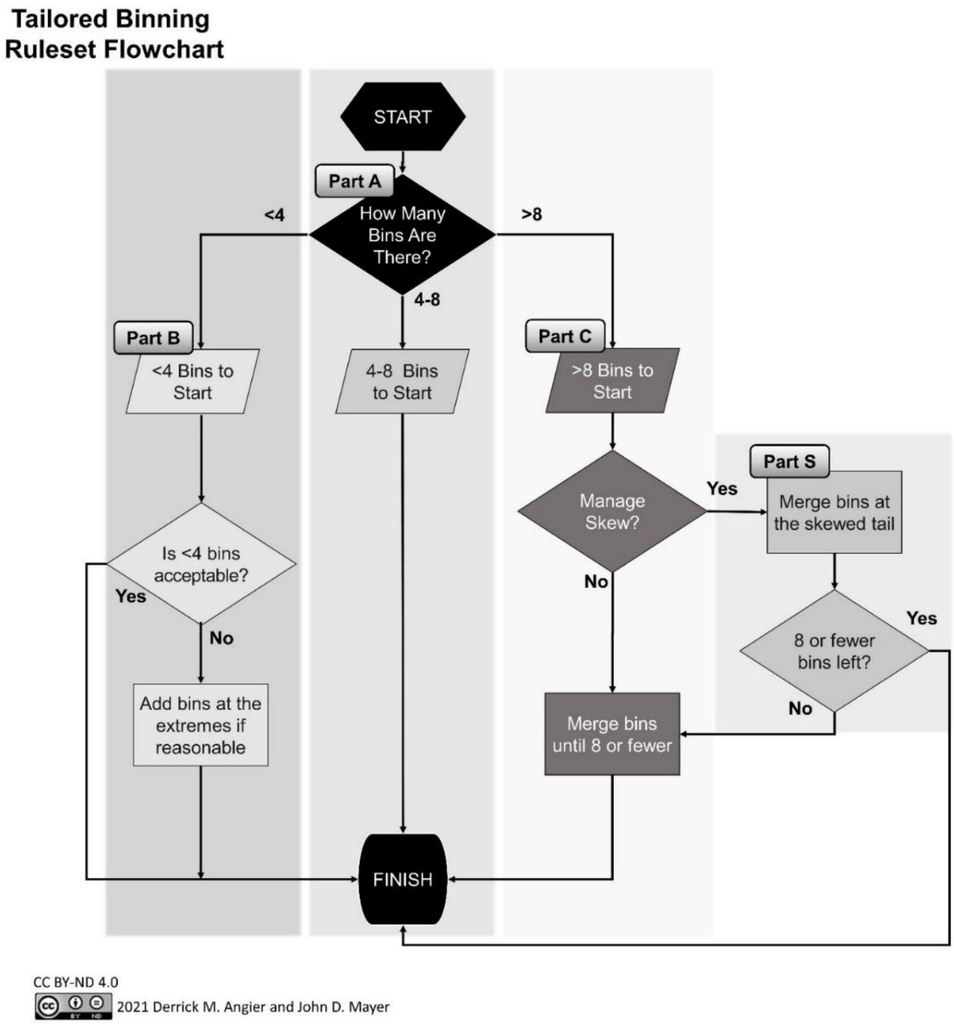

Figure 1. Tailored Binning Ruleset Flowchart. The approach depicted represents the formal rules applied and examined here. A relaxed form of tailored binning used in subsequent papers is described in the General Discussion.

Upon reaching the criterion (in this case, four to eight bins) binning is complete for a given item (Figure 1, "Finish"). The Technical Supplement, Appendix A, provides a step-by-step description of the procedure (Angier & Mayer, 2025). The procedure can also be streamlined by creating preset templates, as described in the Discussion. In this study, however, we elected to follow a standardized, rule-based procedure for purposes of comparison.

4. Study 1 Results

4.1 The Data Set Prior to Binning

Across the 52 items, none had fewer than four different response values (i.e., four starting bins), 36 items had more than eight starting bins, while 16 items had between four and eight, inclusive, starting bins.

4.2 Analyses Used for Comparisons Across Methods

Statistical analyses compared the number and proportion of populated and empty bins as well as the item skewness across the three binning methods. To do so, we either used a one-way ANOVA or a Welch one-way ANOVA. We used the latter when comparison distributions were either highly skewed or lacked homogeneity of variance as this test accommodates skewed data up to levels of skewness of 2.0 and is especially resistant to Type 1 errors when comparing distributions with large differences in skewness (Delacre et al., 2019; Myers et al., 2013); or a Kruskal-Wallis H mean rank test, a non-parametric test, for when one or more comparison distributions were extremely non-normal. For analyses comparing the number of items that contained at least one empty, z proportion tests were used.

4.3 Number of Bins Generated for Each Method

Both Tailored and Equal Interval binning methods led to 100% of items meeting the four-to-eight criterion. Equal Percentage binning met this criterion for just 38 of 52 items (73.1%), a significant difference in proportion compared with Tailored and Equal Interval methods ($z_{prop} = 4.02$, $p < .001$, two-tailed).

Table 1. Comparisons Among Binning Methods for Study 1 Across the 52 Items

	Tailored Bins		**Equal Percentage Bins**		**Equal Interval Bins**		**Sign. of Diff Across Methods**		
	M	***SD***	***M***	***SD***	***M***	***SD***			
Skewness in Binned Item Responses Across Methods (Average Across Items)									
Skewness	1.73	1.15	1.22	.72	5.65	4.05	$_{Welch's}F(2,86.73) = 31.85$***		
Populated, Empty, and Total Bins per Item (Count)									
Populated	7.23	1.06	4.48	1.46	5.48	1.57	$_{Welch's}F(2,98.71) = 65.16$***		
Empty[a]	0	0	0	0	1.98	1.85	χ^2 87.85		
Total bins[a]	7.23	1.06	4.48	1.46	7.46	1.09	$\chi^2(2) = 83.10$***		
Proportion of Populated and Empty Bins per Item									
Populated[a]	0	0	0	0	.25	.23	$\chi^2(2) = 87.85$***		
Empty[a]	1.00	0	1.00	0	.75	.23	(same)		
Number of Items (of 52) That Contain One or More Empty Bins									
	Count	**Perc.**	**Count**	**Perc.**	**Count**	**Perc**	*TB v. EP.* [b]	*TB v. EI*	*EP. v. EI.*
	0	0%	0	0%	35	67.3%	$Z_p = 0.00$	$Z_p = -7.26$***	$Z_p = -7.26$***

Notes. *$p < .05$, **$p < .01$, ***$p < .001$. [a.] A Kruskal-Wallis H mean rank test was used to compare across methods. [b.] Difference between methods in the proportion of items that contain empties measured with (Z_p) proportion z tests. TB indicates Tailored binning, EP indicates Equal Percentage binning, and EI indicates Equal Interval binning.

The total number of bins across all items for the three methods were: Equal Interval, at $M = 7.46$, Tailored binning, $M = 7.23$, and Equal Percentage, $M = 4.48$ (Table 1, "Total bins" row). The number of total bins was significantly different across methods ($\chi^2(2) = 83.10$ $p < .001$; see Table 1, "Total bins" row). Both the Equal Interval and Tailored approaches carried through more information (as indexed by bins) than Equal Percentage. The number of total bins was used in the hypotheses below to calculate the proportions of populated and empty bins.

4.4 Tests of Study 1 Hypotheses

Did Tailored Binning Result in a Greater Reduction of Skewness in Category Distributions than the Other Methods? (Hypothesis 1)

Hypothesis 1 was that Tailored binning would reduce item skewness more than the other two methods. Results showed that the skewness differed across the three binning methods, Welch's $F(2, 86.73) = 31.85$, $p <.001$. A

Tukey HSD test showed that Tailored binning and Equal Percentage did not significantly differ in their skewness (p = .549), but Tailored binning and Equal Percentage each had significantly lower skewness than Equal Interval (p < .001), (see Table 1, "Skewness in Binned Item Responses").

Did Tailored Binning Produce More Populated Bins Per Item than the Alternatives? (Hypothesis 2)

Methods that yield more populated bins may be more informative relative to having fewer, or more sparsely populated, bins. Tailored binning had the highest average number of populated bins (i.e., bins that contain at least one response) per item (M = 7.23 out of a maximum possible of 8), followed by Equal Interval (M = 5.48), and Equal Percentage (M = 4.48); Welch's $F(2, 99.71) = 65.16$, p <.001. A Tukey HSD test showed that each binning method was different from the others in the number of created populated categories (p <.001; See Table 1 "Populated" row). The proportion of bins that were populated out of the total number of bins, per item, was approximately consistent with the prior comparison of the number of populated bins per item, with significant differences between the binning methods $\chi^2(2) = 87.85$, p < .001. See Table 1 for both the number and proportion of populated bins for all three methods.

Follow-up

We also tested the number and proportion of empties per item across methods for an alternative analysis of the same data (Table 1, "Empty" row). The Kruskal-Wallis H tests indicated that Equal Interval had a significantly higher number of empty bins relative to Tailored and Equal Percentage. Regarding proportions, results showed the same rank order, but inverse, with Tailored binning and Equal Percentage showing no empty bins and Equal Interval having the inverse of its proportion of populated bins.

Did Tailored Binning Create a Smaller Number of the 52 Items with Empty Categories than the Other Methods? (Hypothesis 3)

We tested whether Tailored binning was less likely to result in items that contained one or more empty categories. The 52 items exhibited no empty categories in either Tailored or Equal Percentage binning, but 35 (of 52) had empty bins using the Equal Interval binning method. The proportion of 35 items with empty bins (67.3%) was significantly different from zero (see Table 1, bottom row).

5. Study 1 Discussion

In Study 1, we applied the three different binning procedures to our data following a strict form of Tailored binning (a more relaxed form is described in the Discussion) and examined the results according to several criteria. Both Tailored and Equal Interval yielded a desirable number of bins—between four to eight across items. The Equal Percentage method met this criterion for 38 of the 52 items (73.1%), with 14 items exhibiting between one and three bins.

Tailored and Equal Percentage did not significantly differ in their skewness (1.73 and 1.22), and both were superior in terms of reducing skewness relative to Equal Interval binning (Skewness of 5.65).

Tailored binning led the three methods in the number of populated bins, followed by Equal Interval, and then Equal Percentage. Both Tailored and Equal Percentage had 100% of their bins populated across all items (the ruleset for Tailored binning attempts to prevent non-populated bins). By comparison, Equal Interval yielded 75% populated bins. The number of empty bins, the inverse of the populated bins was, similarly, 0% for Tailored and Equal Percentage methods and 25% for Equal Interval.

Tailored binning performed either the best or competitively when looking at: the mean total of bins, the number of items that met the four-to-eight criterion, item skewness, the number and proportion of populated bins, and the number of items that did not contain an empty.

Having compared the results of three binning methods on our data according to several criteria and finding significant differences between them, we wondered how well the bins developed in Study 1would map onto data from the same survey completed by a new sample. That is, would the bins created by the three methods in Study 1 function in similar ways on a new sample.

6. Study 2

A researcher might reasonably ask whether the bins created in Study 1 would be sufficiently stable across samples to use as the basis for future research. To see whether the bins created by the Tailored, Equal Percentage, and Equal Interval methods in Study 1 would exhibit similar characteristics when applied to another sample, we examined results from a second sample, also with open-responding items, in Study 2.The dataset is also part of a broader set of studies reported in Angier (2025).

6.1 Study 2 Hypotheses

In Study 2, we tested:

Hypothesis 1. All three binning methods would reduce the skewness of data distributions in the same rank order and to a degree roughly equal to Study 1. We anticipated that Equal Percentage will have the most reduced skewness, followed by Tailored binning and then Equal Interval. As before, we calculated the skewness per item for each method.

Hypothesis 2. The number and proportion of populated bins per item would be consistent across studies and their samples, and conversely for empty bins. We anticipated that Tailored binning will have the largest number of populated bins, followed by Equal Interval and then Equal Percentage. Equal Interval is expected to contain a much higher number of empties per item relative to the other methods.

Hypothesis 3. The binning methods would be consistent across studies regarding the number of items that contain an empty bin. Tailored binning and Equal Percentage would exhibit a smaller proportion of items with any empty bins than Equal Interval, as found in Study 1.

7. Study 2 Method

7.1 Participants

The Study 2 sample consisted of 246 undergraduate students (85 male; 158 female; 3 other) with a mean age of 19.53 from the University of New Hampshire who took the same survey as in Study 1 approximately six months after the original study. Students were recruited through the UNH Psychology SONA study recruitment system. As with Study 1, screening is reported in detail elsewhere (Angier, 2021), and resulted in the removal of 43 original participants for such issues as overly quick responding or excess omitted data.

7.2 Methods and Measure

The same lifespace items used in Study 1 were administered to the Study 2 sample. As in Study 1, all 246 respondents of this lifespace survey were used in the binning and analyses. As the survey is a branching survey, 113 participants answered just 41 of the 52 items. We report data for all the items, acknowledging that the N dropped from 246 to 133 for 11 items.

Binning Procedures

The bins created for Study 1 were newly applied, without change, to the open-responding data of Study 2.

8. Study 2 Results

8.1 General Considerations Regarding the Statistical Analyses

All statistical analyses comparing the number or proportion of populated and empty bins, and the item skewness across the three binning methods used the same tests as the analyses employed in Study 1, with a few additions for analyses comparing Study 1 with Study 2. For comparing the skewness per item and the number of populated bins per item each method generated across studies, independent t-tests were used. When comparing the number of empty bins per item across studies for each method, non-parametric Mann-Whitney U tests were used instead of independent t-tests as the distributions were non-normal due to the complete lack of empty bins for some methods. Chi-square tests of association were used when comparing the number of items that contain at least one empty across studies for each method. Note that the number of total bins does not change across studies within method, therefore we compared the number of populated and empty bins between Study 1 and Study 2 for each method, but not the proportional differences across studies, which would now be redundant.

8.2 Tests of Study 2 Hypotheses

Did the Average Skewness Across Binning Methods Remain the Same Across Studies for Each of the Binning Methods? (Hypothesis 1)

Average item skewness did not differ significantly across Studies 1 and 2 for any of the three binning methods: Tailored binning, $t(102) = -.74$, $p = .462$; Equal Percentage, $t(102) = -1.00$, $p = .270$; Equal Interval, $t(102) = 1.33$, $p = .900$ (see Tables 1 and 2 for means and standard deviations for skewness per item).

Did the binning methods have the same rank order regarding skewness?

The three methods were again significantly different in Study 2 in item skewness (Welch's $F(2, 88.03) = 32.92$, p <.001). A Tukey HSD test showed that Tailored binning and Equal Percentage each had significantly reduced per-item skewness relative to Equal Interval (see Table 2), but Tailored binning and Equal Percentage did not significantly differ from each other, maintaining the same rank order as in Study 1.

Did the Number of Populated Bins per Item Remain the Same for Each Method-of-Binning When Applying the Bins Created in Study 1 to Study 2? (Hypothesis 2)

To test whether the binning methods would show similar numbers of populated bins in Study 2 as in Study 1, we conducted independent t-tests comparing each method's number of populated bins per item, from Study 1 to 2. The number of populated bins per item across Study 1 and Study 2 showed minimal change that did not reach significance for Equal Interval, $t(102) = 0.14$, $p = .889$ and Tailored binning, $t(102) = 1.05$, $p = .295$. Equal Percentage binning showed no numerical change in the number of populated bins across studies in any items (see Tables 1 and 2, "Populated" row).

Regarding the inverse—the number of empty bins per item— Mann-Whitney U tests found no statistically significant changes between Studies 1 and 2 for Equal Percentage (no difference) and Equal Interval, $U = 1349.00$, $p = .984$, while Tailored binning exhibited more empty bins per item compared to Study 1, $U = 1144.00$, $p = .003$ (see Tables 1 and 2, "Empty" row).

Table 2. Comparisons Among Binning Methods for Study 2 Across the 52 Items

	Tailored Bins		Equal Percentage Bins		Equal Interval Bins		Sign. of Diff Across Methods		
	M	***SD***	***M***	***SD***	***M***	***SD***			
Skewness in Binned Item Responses Across Methods (Average Across Items)									
Skewness	1.92	1.40	1.38	.89	4.85	2.95	$_{Welch's}F(2,88.03) = 32.92$***		
Populated, Empty, and Total Bins per Item (Count)									
Populated	7.00	1.17	4.48	1.46	5.35	1.61	$F(2,153) = 41.95$***		
Empty[a]	.24	.58	0	0	2.12	1.91	$\chi^2(2) = 73.86$***		
Total bins[a]	7.23	1.06	4.48	1.46	7.46	1.09	$\chi^2(2) = 83.10$***		
Proportion of Populated and Empty Bins per Item									
Populated[a]	.97	.08	1.00	0	.74	.24	$\chi^2(2) = 72.88$***		
Empty[a]	.03	.08	0	0	.27	.24	$\chi^2(2) = 70.57$***		
Number of Items (of 52) That Contain One or More Empty Bins									
	Count	**Perc.**	**Count**	**Perc.**	**Count**	**Perc**	*TB v. EP.*[b]	*TB v. EI*	*EP. v. EI.*
	8	15.4%	0	0%	37	71.2%	$Z_p = 2.94$**	$Z_p = -5.74$***	$Z_p = -7.58$***
Difference[c]	8**	15.4%**	0	0%	2	3.9 %			

Notes. *$p < .05$, **$p < .01$, ***$p < .001$. [a] A Kruskal-Wallis H mean rank test was used to compare across methods. [b] Difference between methods in the proportion of items that contain empties measured with (Z_p) proportion z tests. [c] Difference from Study 1 to Study 2 measured with Chi-square tests of association. TB: Tailored binning, EP: Equal Percentage binning, and EI: Equal Interval binning.

Did the three binning methods exhibit the same rank order regarding populated bins?

The three binning methods differed in the number of populated bins $F(2, 153) = 41.95$, $p <.001$, and a Tukey HSD test showed that each binning method was different from the other ($p \leq .001$). Indeed, the rank-order of the three methods was the same as in Study 1. Tailored binning had the highest number of mean populated bins (Study 2 $M = 7.00$, $SD = 1.17$; Study 1 $M = 7.23$, $SD = 1.06$), followed by Equal Interval (Study 2 $M = 5.35$, $SD = 1.61$; Study 1 $M = 5.48$, $SD = 1.57$), and Equal Percentage had the lowest number of mean populated bins with no change from Study 1 (Study 2 $M = 4.48$, $SD = 1.46$; Study 1 $M = 4.48$, $SD = 1.46$).

The proportion of total bins that were populated across the three methods also differed in the same rank order as in Study 1, with Tailored binning populating .97 of bins on average, Equal Percentage at 1.00, and Equal Interval at .74, ($\chi^2(2) = 72.88$, Kruskal-Wallace test, see Table 2).

The number and proportion of empties per item were the inverse of the above, and the relationships across methods were the same as when comparing populated bins (see Table 2, "Empty" row).

Was the Number of Items with Empty Bins the Same for the Binning Methods Across Studies 1 and 2? (Hypothesis 3)

Tailored binning showed a statistically significant increase in the number of items with empty bins across samples from zero to eight items from Study 1 to 2 (see Table 2, "Difference" row), while Equal Percentage showed no change, and Equal Interval showed a change of 35 to 37 items, which did not reach statistical significance (see Table 2, bottom rows).

Did the binning methods have the same rank order regarding the proportion of items with empty bins?

Tailored binning resulted in a significantly higher proportion of empty bins than Equal Percentage in Study 2, whereas it had not in Study 1 (see Table 2, bottom right). In Study 2, Equal Percentage had the fewest items with empties (zero), followed by Tailored (eight), and Equal Interval had the most items with empties (37).

9. Study 2 Discussion

Study 2 indicated the stability of findings in a new sample in which the performance of Tailored, Equal Percentage, and Equal Interval binning approaches performed much the same as in Study 1. Because of the similarities across studies, we proceed directly to the General Discussion.

10. General Discussion

Information about binning open-ended potentially non-normally distributed survey responses is relatively scarce in the literature and often focuses on Equal Interval or Equal Percentage approaches (Dash et al., 2011; Knuth, 2013). As the names suggest, Equal Interval binning divides participants' numerical responses into categories of equal range; Equal Percentage creates bins that hold equal numbers of responses (so much as is possible). We compared those to a new, Tailored binning approach described here (see Method, Study 1). That approach encourages researchers to perform a series of specified judgments regarding the possible bins entailed by each item, to enhance the item's bins relative to their meaningfulness and statistical performance.

The purpose of this paper was to explore how Tailored binning compares with the alternatives and if it is beneficial for data management and measure development. We specified several criteria along which to compare the methods: That a given method should (a) yield between 4 and 8 categories, (b) employ mostly populated bins across all its items, (c) exhibit populated bins within items, and (d) moderate skewed responses as much as possible.

In Study 1, we examined the three binning methods' capabilities to meet our criteria for this data. Using the bins from Study 1, we gauged how consistent their performance was with a new sample.

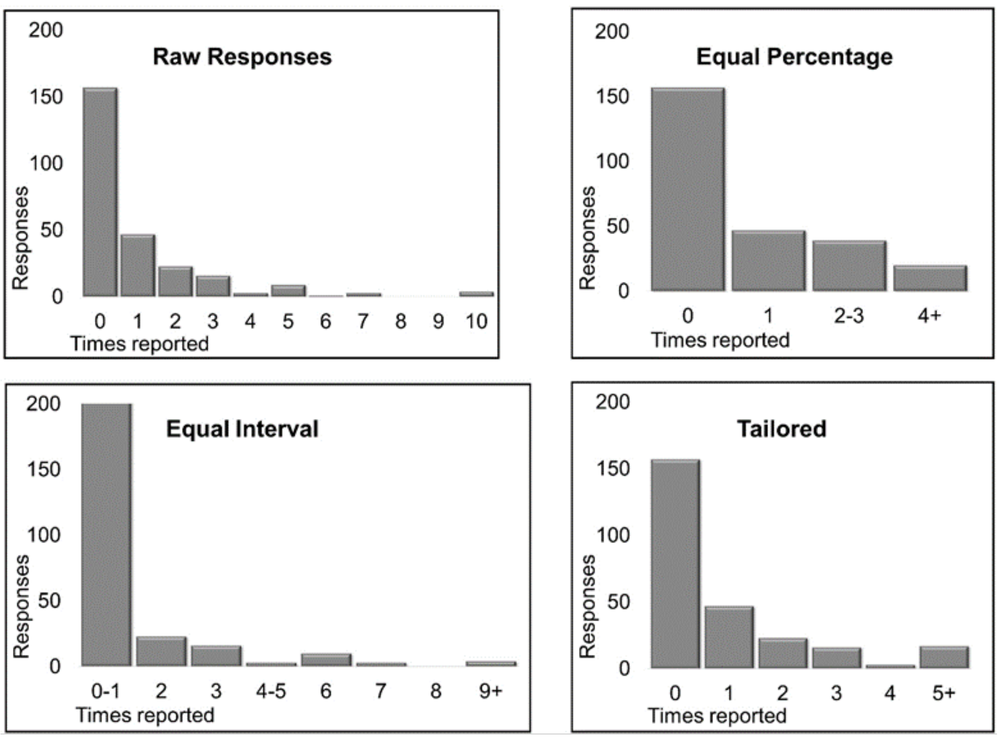

Figure 2. Transformations of Responses to a Sample Item. The figure shows the histograms for raw responses and the results of binning procedures on the item concerning the times a day a gamer had "*abruptly stop playing a video game out of frustration ('ragequit')*" Lifespace items can vary significantly from each other, but this item was chosen as an example as it was relatively average and demonstrated the generally noticeable patterns in binning results.

10.1 Relative Merits of the Approaches

The three binning methods all show that they can reduce the range of responses into organized categories and all three methods exhibited a certain amount of consistency across samples. That said, the Equal Interval binning method was the weakest option, judged by our criteria. Although it exhibited (a) a moderate number of populated bins (fewer than Tailored, more than Equal Percentage), it also produced (b) a higher number of items with empties, (c) higher scale skewness relative to other methods (though the other methods remained statistically skewed), and, observationally, the nature of its binning tended to condense lower response values (e.g., 0, 1, or 2), removing information granularity in skewed data.

A further qualitative advantage of Tailored binning is that researcher judgment may be used for creating binning criteria rules. For example, researchers could mandate that each relevant item has a bin solely for the value of zero; a rule used in the tailored procedures of these studies. Tailored binning allows researchers to use their judgment in key instances to adapt the binning rules according to their goals, potentially enhancing the development of closed-ended item response alternatives and improving item-by-item comparisons.

An example of binning for a single, representative item is in Figure 2. The item asked about “ragequitting”—abruptly quitting a videogame out of frustration or anger with the game. The raw responses (top left) ranged from zero to 10 average rage quits a day over the prior week, with the majority responding with zero. Equal Interval (bottom left) combined the two most common responses, 0 times and 1 time, to a single interval, leading to a more skewed distribution than the original and condensing responses that might be qualitatively different (never ragequitting might be different from any incident of a ragequit). The same method also introduced an empty bin for responses of eight. Equal Percentage (top right), on the other hand, created only four bins but, in this case at least, more meaningfully represented the data. Although Equal Percentage provided effective organization, it limited the range of observable responses relative to the other methods (e.g., responses of two and three combined in a bin, and “four or more” as the highest value bin). Tailored binning led to a binning distribution similar to Equal Percentage but with a higher number of bins, six, representing a broader range of potential response values and displaying more differentiaion at the high end of the distribution. Tailored binning (bottom right) had a higher response ceiling and, unlike Equal Percentage, provided two separate bins for responses of two and three, thus preserving more of the observable differences in response values. Increased bin differentiation, including among the higher value bins, may indicate meaningful differences in respondent’s reported behaviors, and in this case may represent significant differences in frustration level and self-control when gaming. The information communicated by Tailored binning may potentially enhance later comparisons and relationships between the item and other criteria.

10.2 Limitations

Limits to a Single Scale

Despite the strengths of applying the techniques to real-life data collected in two waves, this study was limited in several respects. For one, only one lifespace scale was used, and there might be specific variance associated with its 52 items such that other scales might behave differently. In addition, the two independently collected samples employed shared the same, specific overall demographics. While advantageous for the purposes of measure development and evaluating the reliability of binning methods, further study of different and diverse samples would be beneficial in assessing the generalizability of these binning approaches. The criteria set by the authors will not necessarily be the same criteria required and set by other researchers (e.g., setting four to eight bins as desirable), so further research is needed to confirm the performance of the binning methods when applied to the potential variations in binning criteria.

Time Investment and the use of Streamlined Tailored Binning

Another limitation is the time investment necessary to go throught the steps outlined here for Tailored binning in their entirety. To address that limitation, two possibilities suggest themselves for future work: First, the development of computer code to bin the materials using the rules supplied here. The second approach, taken subsequently in our lab, has been the use of Templated Tailored Binning, described next.

10.3 Templated Tailored Binning

As noted in the section on limitations, the Tailored binning procedure here is potentially time consuming if the number of items on a scale is large. In a modifcation of the procedure we refer to as “Templated Tailored Binning”, the tailored technique has been streamlined. To do so, subsequent researchers created several binning templates by grouping items based on shared response patterns such as whether a given behavior appeared to be common multiple times a day, once a day or less, several times a week, and other alternatives. This Templated Tailored

Binning approach is a promising further development of Tailored binning that reduces the time invested by researchers while retaining their common-sense judgments, and involves eyeballing initial data distributions and then matching items to pre-set bins. Further descriptions of the approach can be found in Bryan (2023) and in a lifespace study by Mayer, Caruso, and Panter, with details on the Templated Tailored Binning in the online open-source technical supplement (Mayer et al., 2023, 2024).

10.4 Conclusions

Researchers, particularly in the social sciences, often choose to bin responses to open-ended questions in their research. The process is a foundational step in data analysis upon which subsequent analyses depend, especially when handling lifespace data or similar. A step that, perhaps, has not been as fully developed as might be possible. This exploration of Tailored binning has established a foundation for why a Tailored binning approach may be effective and how it may be successfully carried out. We are excited by the possibilities of Tailored binning for curating and preserving the meaning of participant responses relative to other approaches, as well as the potential variations of the method moving forward such as the tailored template method. Also, future automation of part of the Tailored binning process with statistical software could allow researchers to focus primarily on setting up the procedure and inputting the desired criteria. We are hopeful this investigation into binning improves research measurement in the social sciences and beyond. Measurement is a key aspect of all scientific advancement (Strevens, 2020), and enhancing measurement-related techniques and options will only strengthen the relevant fields of research.

References

Angier, D. M. (2021). Gaming Engagement and Outcomes: The Lifespace of Gamers [Unpublished master's thesis].

Angier, D. M. (2025). Innocent Fun or Unhealthy Habit: What's Your Game? Personality and Engagement with Video Games [Manuscript in preparation].

Angier, D. M., & Mayer, J. D. (2025). Technical Supplement for "A Comparison of Methods for Binning Responses to Open- Ended Survey Items About Everyday Events: A New Tailored-Binning Approach." https://osf.io/v5ty6

Borgers, N., Hox, J., & Sikkel, D. (2004). Response Effects in Surveys on Children and Adolescents: The Effect of Number of Response Options, Negative Wording, and Neutral Mid-Point. *Quality & Quantity, 38*, 17–33. https://doi.org/10.1023/B:QUQU.0000013236.29205.a6

Brackett, M. A., & Mayer, J. D. (2006). The life space: A framework and method to describe the individual's external traits. *Imagination, Cognition and Personality, 26*(1–2), 3–41. https://doi.org/10.2190/8380-1676-H338-N217

Bryan, V. (2023). Knowing Me, Knowing You: The Role of Personal Intelligence in Romantic Relationship Conflict Communication. Doctoral Dissertations. https://scholars.unh.edu/dissertation/2774

Buss, D. M., & Craik, K. H. (1983). The Act Frequency Approach to Personality. *Psychological Review, 90*(2), 105–126. https://doi.org/10.1037/0033-295X.90.2.105

Chapman, B. P., & Goldberg, L. R. (2017). Act-frequency signatures of the Big Five. *Personality and Individual Differences, 116*, 201–205. https://doi.org/10.1016/j.paid.2017.04.049

Current Population Survey: 2019 ASEC Technical Documentation. (2020). Washington: U.S. Census Bureau. Retrieved from https://www2.census.gov/programs-surveys/cps/techdocs/cpsmar19.pdf

Dash, R., Paramguru, R. L., & Dash, R. (2011). Comparative Analysis of Supervised and Unsupervised Discretization Techniques. *International Journal of Advances in Science and Technology, 2*(3), 29–37.

De Brouwer, P. (2021). Data Binning. In The Big R-Book: From Data Science to Learning Machines and Big Data (1st ed., pp. 343–361). Wiley. https://doi.org/10.1002/9781119632757

Delacre, M., Leys, C., Mora, Y. L., & Lakens, D. (2019). Taking Parametric Assumptions Seriously: Arguments for the Use of Welch's F-test instead of the Classical F-test in One-Way ANOVA. *International Review of Social Psychology, 32*(1), 13. https://doi.org/10.5334/irsp.198

Doane, D. P. (1976). Aesthetic Frequency Classifications. *The American Statistician, 30*(4), 181–183. JSTOR. https://doi.org/10.2307/2683757

Freedman, D., & Diaconis, P. (1981). On the histogram as a density estimator:L2 theory. *Zeitschrift Für Wahrscheinlichkeitstheorie Und Verwandte Gebiete, 57*(4), 453–476. https://doi.org/10.1007/BF01025868

Ivcevic, Z. (2007). Artistic and Everyday Creativity: An Act-Frequency Approach. *The Journal of Creative Behavior, 41*(4), 271–290. https://doi.org/10.1002/j.2162-6057.2007.tb01074.x

Ivcevic, Z., & Mayer, J. D. (2009). Mapping dimensions of creativity in the life-space. *Creativity Research Journal, 21*(2–3), 152–165. https://doi.org/10.1080/10400410902855259

Knuth, K. H. (2013). Optimal Data-Based Binning for Histograms. arXiv:Physics. http://arxiv.org/abs/physics/0605197v2

Kotsiantis, S., & Kanellopoulos, D. (2006). Discretization Techniques: A recent survey. *GESTS International Transactions on Computer Science and Engineering, 32*(1), 47–58.

Lavielle, M., & Bleakley, K. (2011). Automatic data binning for improved visual diagnosis of pharmacometric models. *Journal of Pharmacokinetics and Pharmacodynamics, 38*(6), 861–871. https://doi.org/10.1007/s10928-011-9223-3

Mael, F. A. (1991). A Conceptual Rationale for the Domain and Attributes of Biodata. *Personnel Psychology, 44*(4), 763–792. https://doi.org/10.1111/j.1744-6570.1991.tb00698.x

Mael, F. A., & Hirsch, A. C. (1993). Rainforest Empiricism and Quasi'-Rationality: Two Approaches to Objective Biodata. *Personnel Psychology, 46*(4), 719–738. https://doi.org/10.1111/j.1744-6570.1993.tb01566.x

Mayer, J. D., & Bryan, V. M. (2024). On Personality Measures and Their Data: A Classification of Measurement Approaches and Their Recommended Uses. *Personality and Social Psychology Review, 28*(3), 325-345. https://doi.org/10.1177/10888683231222519

Mayer, J. D., Carlsmith, K. M., & Chabot, H. F. (1998). Describing the Person's External Environment: Conceptualizing and Measuring the Life Space. *Journal of Research in Personality, 32*(3), 253–296. https://doi.org/10.1006/jrpe.1998.2220

Mayer, J. D., Caruso, D. R., & Panter, A. T. (2023). Personal Intelligence Lifespace Inventory (PILSI) Combined Technical Report. Retrieved from https://osf.io/3ea4r/

Mayer, J. D., Caruso, D. R., & Panter, A. T. (2024). Lifespace Patterns of College Students High and Low in Personal Intelligence. Collabra: Psychology. https://doi.org/10.1525/collabra.90222

Myers, J. L., Well, A. D., & Lorch, R. F. Jr. (2013). Research Design and Statistical Analysis: Third Edition. Routledge.

Paunonen, S. V. (2003). Big Five factors of personality and replicated predictions of behavior. *Journal of Personality and Social Psychology, 84*(2), 411–424. https://doi.org/10.1037/0022-3514.84.2.411

Perkey, H., Sinclair, S. J., Blais, M., Stein, M. B., Neal, P., Pierson, A. D., & Slavin-Mulford, J. (2018). External validity of the Psychiatric Diagnostic Screening Questionnaire (PDSQ) in a clinical sample. *Psychiatry Research, 261*, 14–20. https://doi.org/10.1016/j.psychres.2017.12.011

Prieto Curiel, R., Cabrera Arnau, C., Torres Pinedo, M., González Ramírez, H., & Bishop, S. R. (2020). A continuous binning for discrete, sparse and concentrated observations. *MethodsX, 7*, 100709. https://doi.org/10.1016/j.mex.2019.10.020

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika, 66*(3), 605–610. https://doi.org/10.1093/biomet/66.3.605

Semega, J., Kollar, M., Creamer, J., & Mohanty, A. (2020). Income and Poverty in the United States: 2018. Retrieved from https://www.census.gov/content/dam/Census/library/publications/2019/demo/p60-266.pdf

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the number of response options matter? Psychometric perspectives using personality questionnaire data. *Psychological Assessment, 31*(4), 557–566. https://doi.org/10.1037/pas0000648

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Strevens, M. (2020). The Knowledge Machine: How Irrationality Created Modern Science. Liveright.

Sturges, H. A. (1926). The Choice of a Class Interval. *Journal of the American Statistical Association, 21*(153), 65–66. https://doi.org/10.1080/01621459.1926.10502161

Sulastri, A., Handoko, M., & Janssens, J. M. A. M. (2015). Grade point average and biographical data in personal resumes: Predictors of finding employment. *International Journal of Adolescence and Youth, 20*(3), 306–316. https://doi.org/10.1080/02673843.2014.996236

Tukey, J. W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics, 33*(1), 1–67. https://doi.org/10.1214/aoms/1177704711

Copyrights