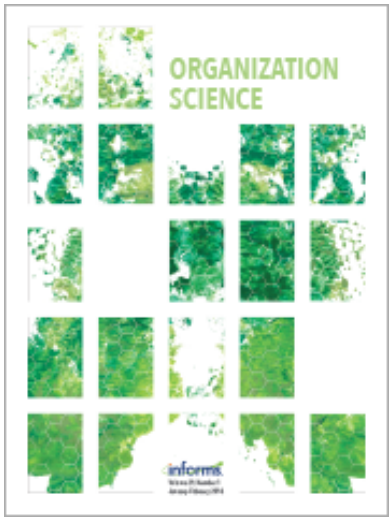


Organization Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Misfit Bias

Bryan K. Stroube, Keyvan Vakili, Michaël Bikard

To cite this article:
Bryan K. Stroube, Keyvan Vakili, Michaël Bikard (2025) The Misfit Bias. Organization Science 36(5):1676-1689.
https://doi.org/10.1287/orsc.2023.17462

https://pubsonline.informs.org/journal/orsc

ORGANIZATION SCIENCE
Vol. 36, No. 5, September–October 2025, pp. 1676–1689
ISSN 1047-7039 (print), ISSN 1526-5455 (online)

The Misfit Bias

Bryan K. Stroube,[a,*] Keyvan Vakili,[a] Michaël Bikard[b]

[a]Strategy and Entrepreneurship Department, London Business School, London NW1 4SA, United Kingdom; [b]Strategy Department, INSEAD, F-77305 Fontainebleau, France

*Corresponding author

Contact: bstroube@london.edu, https://orcid.org/0000-0002-1785-3267 (BKS); kvakili@london.edu, https://orcid.org/0000-0002-9180-9360 (KV); michael.bikard@insead.edu, https://orcid.org/0000-0002-0298-7246 (MB)

Received: March 8, 2023
Revised: July 9, 2024; December 13, 2024
Accepted: January 13, 2025
Published Online in Articles in Advance: March 25, 2025

https://doi.org/10.1287/orsc.2023.17462

Abstract. Consumers and other audiences often penalize products that combine unrelated elements. In this paper, we document the consequences of that penalty for the evaluation of the elements being combined. Building on the idea that audiences cannot fully disentangle the quality of "fit" between elements from the quality of the elements individually, we argue that audiences are likely to direct their dislike of a misfit product to the individual elements being combined. Using an archival study of the music industry and an online experiment with photographic galleries, we find that evaluations of individual elements (songs, photographs) are influenced by product-level fit (albums, galleries). Elements of misfit products are evaluated less favorably than they would have been otherwise. Moreover, this bias is exacerbated when the evaluation of the whole product is emphasized. We discuss the implications of this "misfit bias" for the innovation, entrepreneurship, and categories literatures.

Supplemental Material: The online appendix is available at https://doi.org/10.1287/orsc.2023.17462.

Keywords: innovation • evaluations • recombination • bias • categories • cultural products • creativity

Introduction

New products, services, and ideas often combine distant or unrelated elements. Indeed, such combinations are a central feature of innovation (Schumpeter 1934) despite the fact that, on average, consumers and evaluators tend to discount them (Negro et al. 2010, Boudreau et al. 2016). Scholars have examined this discount using a multitude of theoretical perspectives, including innovative, cognitive, and normative lenses (Cutolo and Ferriani 2024), in a wide range of empirical settings such as feature films (Hsu 2006), restaurants (Rao et al. 2005, Kovács and Hannan 2010), wines (Negro and Leung 2013), loans (Leung and Sharkey 2014), scientific publications (Uzzi et al. 2013), and new technologies (Fleming 2001, Fleming et al. 2007). The core idea is that individuals are more likely to ignore or punish a product with misfit[1] elements because they may view it as confusing or illegitimate (Hsu 2006, Negro and Leung 2013, Leung and Sharkey 2014).[2] Although previous literature has extensively studied how crossing boundaries affects the evaluation of new combinations, it has largely overlooked the impact on the underlying elements being combined.

One reason for this omission is that in many combinations, the elements being combined are not easily distinguishable. For example, movies often mix elements from different genres, making it difficult to decompose a consumer's rating of a romantic comedy into the independent quality of its romance versus comedy elements. This type of "melting pot" combination fuses together different elements to create something new. By contrast, other combinations are more akin to "salad bowls" in that they contain elements that can be independently identified, consumed, and evaluated. A music album contains songs; a gallery contains artwork; a book contains chapters; an academic course contains individual sessions; and an interior design contains different furniture pieces of various shapes, materials, and colors. Notably, producers of albums, galleries, books, and courses evaluate the quality of individual elements as they choose what to include in their future creations. Misjudgment about the quality of an element because of its inclusion in a novel combination might thereby affect the element's prospects, as well as those of its creators. Yet prior research has largely overlooked the possibility of such an effect.

In this paper, we ask how the quality perception of these individual elements is affected by the combinations in which the elements appear. Specifically, we posit that when faced with a product that combines distant or unrelated elements, consumers will not only react more negatively (on average) to the product itself—the baseline prediction from the literature—but will also evaluate its individual elements more harshly.

Our argument starts with the observation that the quality of any combinatory creation is a function of at least two traits: (1) the individual quality of each element being combined and (2) the extent of fit (or misfit) between those elements. Drawing on research from organizational psychology on the halo effect, we then argue that, although these two traits, namely fit and the quality of each element, are in principle independent from each other, individuals may not fully distinguish them from each other as part of their evaluation.

The halo effect, first identified by Thorndike (1920), highlights how evaluators often fail to assess distinct attributes independently (Nisbett and Wilson 1977). For example, research shows that consumers infer health benefits for unspecified aspects of a product based on claims about other features (Andrews et al. 1998, Roe et al. 1999), and physically attractive individuals are often rated more favorably on unrelated traits like intelligence or sociability (Feingold 1992). The theoretical models of the halo effect suggest that the evaluation of an attribute can bias the evaluation of other attributes either through direct spillover or through affecting the general impression of the overall target (person or product) (Fisicaro and Lance 1990; Lance et al. 1994a, b). Given that fit (or misfit) by definition is a general trait of a multielement product, a negative perception of fit can spillover on the evaluation of other traits, namely the quality of individual elements, both directly and via lowering the general impression of the product as a whole. Consequently, we argue that an element of a product will be rated less favorably when the product has lower fit between its elements than when the product has higher fit between its elements. Following the same logic, we further argue that this penalty may be exacerbated when individuals first conduct a holistic evaluation of a product before evaluating its elements, because thinking about the product as a whole prompts individuals to see misfit more saliently.[3]

Testing these arguments empirically is difficult, not least because it requires large-scale performance data on the evaluation of individual elements. In many contexts, such data are not systematically measured; for example, universities, in an effort to help students select between courses, collect *course* evaluations but not individual session evaluations. Moreover, examining the effect of misfit on the evaluation of elements faces identification problems because of potential spurious correlation between the quality of elements and the extent of fit between them in a product.

To address these challenges, we employ a two-part empirical strategy using observational data and an experiment. Our observational analyses leverage features of the music industry, where although songs have historically been released and purchased as part of an album, they also have their own performance outcomes. Our analyses show a strong negative relationship between the level of misfit between songs on an album and an individual song's likelihood of making it to the Billboard Hot 100 chart. We then exploit a quasi-natural experiment in the industry: the rapid change in the consumption of music from being mostly in the form of combinations (albums) to individual elements (songs) in the early 2000s. In line with our arguments, the negative effect of misfit shrinks significantly when the bundle is less salient—that is, after consumers shift from album consumption to the less bundled consumption of individual songs.

The archival analysis has the advantage of showing how misfit affects the evaluation of elements, using real-world data in an important industry. However, this analysis suffers from potential endogeneity and measurement issues. Therefore, to complement the archival analysis, we also conduct an online laboratory experiment in which we fully randomize the extent of fit between elements while controlling for their quality. We find similar results, supporting our arguments.

This paper makes contributions to at least two long-standing literatures. First, a vast literature has highlighted the (generally negative) performance implications of bringing distant elements together for the evaluation of the whole (Hsu 2006, Leung and Sharkey 2014). We seek to complement this stream of work by highlighting that those combinations affect the evaluation not only of the whole but also of its parts. Second, scholars of creativity and innovation have long described distant combinations with optimism. Although such combinations tend to lead to worse average outcomes, they can also produce breakthrough ideas (Chai 2017, Leahey et al. 2017). This paper highlights a hitherto unrecognized cost of distant combinations—that they can negatively affect the evaluations of individual elements. In addition, our theory and findings also speak to a broader set of organizational phenomena related to the difficulty of assessing individual performance in a team or the evaluation of firm resources when they generate value as a bundle.

Empirics

We test our prediction using two different approaches. First, we analyze an archival sample of approximately 473,000 music tracks belonging to about 40,000 albums released between 1998 and 2005. We show (consistent with our first prediction) that the level of misfit between tracks on an album negatively and significantly affects the likelihood that a song from the album will appear on

the Billboard Hot 100 chart, a common indicator of song success. We then show (consistent with our second prediction) that the effect of fit on the success of songs on an album drops significantly after a shift in music consumption from CD albums to unbundled consumption of music tracks after 2001. Although we believe these results are consistent with our arguments, the analysis of observational data introduces some clear limitations, including the fact that we cannot observe the same element as it is used in different combinations, making it impossible to completely rule out some types of omitted variable bias.

Therefore, we also design an online experiment in the context of art galleries to further isolate the role of fit as the driving mechanism. Our results show that even when the quality of an element, in this case a photo, is fixed, its evaluation by individuals significantly drops when it is included in a less-fit gallery compared with a more-fit gallery. We also find that the misfit bias increases when participants are asked to rate the gallery before the individual photos (as opposed to rating the photos before the gallery).

Archival Analysis in the Music Industry

In our first set of analyses, we examine how the extent of misfit between tracks on a music album influences the success rate of individual tracks. In line with our theoretical prediction, we expect that tracks included on a less-fit album will face a perceptual penalty and therefore perform worse in the market than tracks on albums with greater fit. Furthermore, as albums become less important as units of consumption, this penalty will decrease.

To test the first argument, we first construct a measure of misfit at the album level using the sonic features of songs included on the album and subsequently show that songs on more misfit albums are less likely to appear on the Billboard Hot 100 charts.

To test the second of these arguments, we use an important shock in the music industry: the dramatic shift in music consumption around 2001 from predominantly CD albums to the unbundled consumption of single tracks. Music sales reached their peak in 2000. That year, individual consumers in the United States purchased more than 900 million CD albums (Covert 2013). Singles sales—the sale of individual song tracks—were a drop in the bucket until then, particularly because it was not economically justifiable in most cases for labels and music producers to put a single track on a CD or cassette. As a result, aside from listening to radio stations, until 2000, music consumption was primarily in the form of purchasing and listening to music albums on CDs and cassettes. However, the music digitization trend that began with the introduction of the mp3 format in 1993 increased significantly in 2001 and 2002, until eventually digital sales of single tracks outpaced sales of CD albums by 2006.[4] The introduction of the iTunes music player in January 2001, the iPod in October 2001, BitTorrent—the file-sharing adware used heavily for downloading digital media—in July 2001, and eventually the release of the iTunes store in April 2003 changed the face of the music industry in the span of two years. Figure 1 shows the sudden downfall of album sales beginning in 2000. We use the year 2001 as the beginning of the shock and conduct various robustness checks.

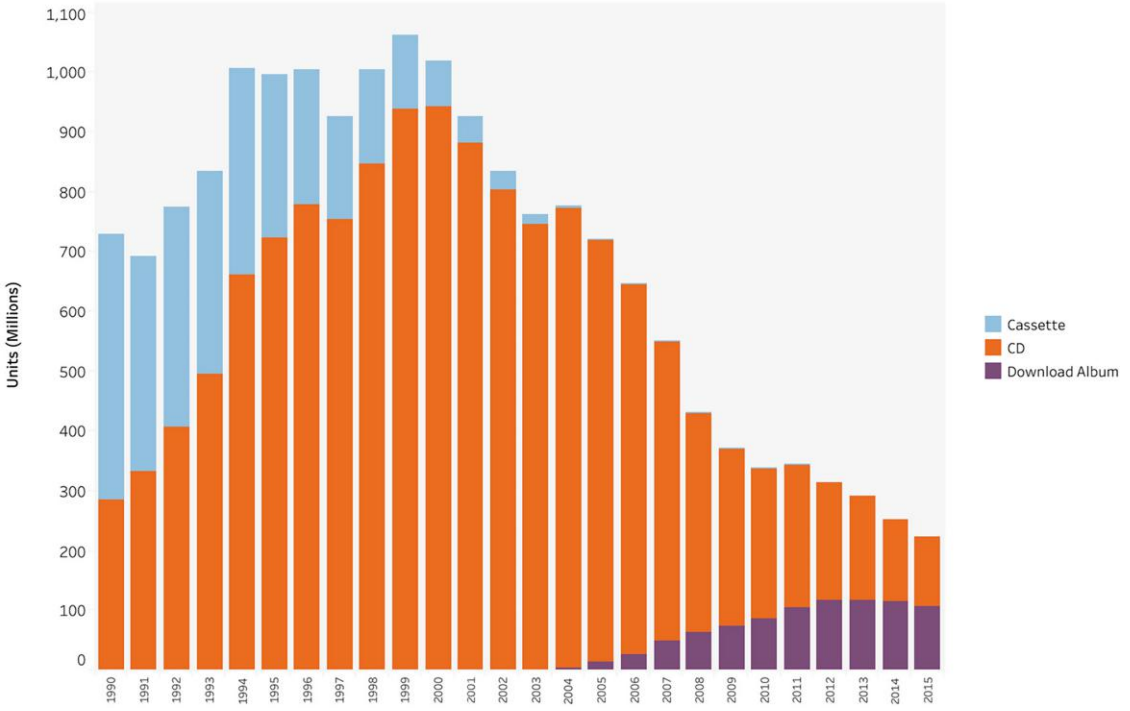
Our theoretical argument suggests that before 2001, when music was consumed primarily in the form of albums, the extent of fit between album tracks should affect the perceived quality and therefore success of these tracks. However, with the shift to unbundled consumption of tracks, the fit between album tracks should matter less for the success of individual tracks. Our core assumption is that the change in consumption patterns that began in 2001 did not affect the level of fit in albums in the short term. Indeed, as we show further below, the average level of fit between album tracks does not materially change during the period we study.

Archival Data

We collected data for our archival analysis from two sources. Data on songs, albums, and artists come from a Spotify data repository available on Kaggle. These data were originally collected via the Spotify API and cover more than 1.2 million songs. For the purpose of our analysis, we used data on albums released between 1998 and 2005. We further excluded albums longer than 148 minutes (equivalent to two standard CDs containing uncompressed stereo digital audio), shorter than 20 minutes, or with fewer than five songs on them. Finally, we excluded songs that were shorter than 1 minute or longer than 9.2 minutes (more than three standard deviation above the mean). We also excluded albums for which we did not have clear artist or genre data. Our final data set includes 31,925 albums and 359,040 tracks.

The data for each track include title, duration, position on the album, the album title, artist names, and release year. In addition, the data include 11 sonic features of each song such as its valence, acousticness, energy, and tempo. These features were derived algorithmically by Echo Nest (acquired by Spotify in 2014) based on the musical content of each song, using advanced music information-retrieval techniques. Together, these features approximately represent what people hear when they listen to a particular track. Following Askin and Mauskapf (2017), we use the features to calculate the distance between songs: in other words, the fit between them. Much research on categories relies on category labels (e.g., "jazz" or "rock" in the context of music) to infer category boundaries and fit. These labels work well when product markets are stable and boundaries between categories are clear. However, in the context of cultural products such as music, they may not provide adequate or accurate information to consumers. As a result,

Figure 1. (Color online) Decline of Album Sales



Source. Recording Industry Association of America; riaa.com/u-s-sales-database/.

individuals may rely more on underlying products' features to make sense of their categories and fit (Lena 2006, Jones et al. 2012, Rossman and Schilke 2014, Askin and Mauskapf 2017). Askin and Mauskapf (2017), in particular, show that a fit measure based on these features can strongly predict the popularity of songs in line with the "optimal distinctiveness" predictions. Each feature is normalized to a zero to one scale to ensure that no specific feature skews the fit measure. Similar to Askin and Mauskapf (2017), we collapsed these 11 measures into a single vector for each song.[5]

Predictor Variables. Our first main independent variable is the level of fit between tracks on an album. We calculate it as the arithmetic mean of cosine similarities between all track-pairs on an album. The formula for calculating the extent of fit between tracks on an album is as follows:

$$Album\_Fit = \sum_{\substack{all\ track\ pairs\ i,j \\ in\ the\ album}} \frac{S_i \cdot S_j}{\|S_i\|\|S_j\|},$$

where S_i and S_j are the vectors of sonic features of tracks i and j of the album, respectively; $S_i \cdot S_j$ is the inner product of the S_i and S_j vectors; and $\|S_i\|$ and $\|S_j\|$ are the sizes of the two vectors. Note that level of fit between tracks on an album is defined at the album level and is the same for all songs on the album.

We illustrate the measure by comparing two well-known albums from our sample: *Kid A*, by Radiohead, released in 2000, and *Come on Over*, by Shania Twain, released in 1997. In his review of the Radiohead album in *Pitchfork*, Brent DiCrescenzo (2000) famously wrote about the diversity of music on *Kid A* (Enis 2020):

> "Comparing this to other albums is like comparing an aquarium to blue construction paper. And not because it's jazz or fusion or ambient or electronic. Classifications don't come to mind once deep inside this expansive, hypnotic world."

David Fricke (2000), in his *Rolling Stone* review of the album, noted, "If you're looking for instant joy and easy definition, you are swimming in the wrong soup". The Wikipedia entry for the album notes that "*Kid A* has been described as a work of electronica, experimental rock, postrock, alternative rock, post-prog, ambient, electronic rock, art rock, and art pop" and cites 13 different external sources to justify those categories (Wikipedia 2022).

Shania Twain's *Come on Over* provides a contrasting example. The album was wildly successful and, according to Guinness World Records, is the all-time "biggest-selling studio album by a female solo artist" (Guinness World Records 2015). Even though Twain, as an artist, pioneered "pop-country" crossover music, the songs on the album did not feature the type of variance that was clear in the Radiohead example. In a review of the album in *Pitchfork*, Allison Hussey (2021) wrote:

> "With the smeared edges of their production, Twain and Lange [the producer] master the illusion of genre, as if they fashioned *Come on Over* into a plastic lenticular print."

Consistent with these descriptions, *Kid A* receives a fit score of 0.68 based on our measure, placing it in the bottom 5% of albums in our sample in terms of fit between its tracks. By contrast, Shania Twain's *Come on Over* receives a fit score of 0.93, placing it in the top 5% of the most-fit albums in our sample.

Dependent Variables. We use the weekly Billboard Hot 100 charts to measure the general performance of songs in our sample. The Billboard Hot 100 is the standard record chart for songs in the United States. The chart ranking is published by *Billboard* magazine each week and incorporates data on sales, radio plays, and, more recently, online streaming in the United States. Appearance on the Billboard Hot 100 is commonly used as a measure of song popularity and performance in research (Askin and Mauskapf 2017, Chang 2023).

We use two different measures to calculate the popularity of songs in our sample. The first variable is a dummy indicator capturing whether a song appeared on the Billboard Hot 100 charts in the year it was released. The second variable captures the number of weeks each song stayed on the chart. We use song titles and release years to match tracks from the Spotify data to the Billboard Hot 100.

We note some caveats with this measure. Theoretically, we are interested in measuring consumers' evaluation of each song. Although the evaluation of a song by consumers certainly affects its sales and radio plays, and hence its likelihood of appearing on the Billboard list, there are at least three issues we wish to address. First, it is possible that how much a musician cares about the Billboard Hot 100 is systematically correlated with her taste for experimentation. For example, musicians who care less about the Billboard list might also do more experimentation and therefore produce albums with higher levels of misfit. Similarly, one could argue that higher-quality artists might produce more fit albums but also spend more on marketing their songs and having them played on the radio. We address these issues, to the extent possible, by including artist fixed effects in our analysis. Artist fixed effects should in principle control for time-invariant qualities of artists and their preference for having their songs on the Billboard list.

Second, one might be concerned that the appearance of a song on the Billboard Hot 100 chart is more affected by what radio stations and retailers like than what the average consumer likes. To the extent that radio stations and retailers can predict what consumers like on average, our measure should still work but contain noise. At the same time, theoretically one could argue that an important audience for any song is the set of individuals deciding whether to play that song on their radio stations or sell it more prominently in their stores. As a result, our arguments would still extend to this group and would naturally lead to the same predictions if we use appearance on the Billboard Hot 100 as a proxy for how the audience (in this case, radio stations and retailers) evaluate a song.

Third, one might be concerned that the 2001 shock to the consumption patterns of consumers has also affected the production side and the broader institutional environment. As a result, what we might attribute as the effect of decline in the salience of album fit among consumers might be instead driven by production-side or institutional factors. There is clearly some evidence that both the production side and the broader institutional environment eventually adapted to the change in consumption patterns (Zanella et al. 2022, Chang 2023). For example, Billboard began tracking and including paid digital downloads in its ranking in February 2005, the final year in our sample. However, in our robustness checks, we show various evidence suggesting that such changes did not occur immediately after the shock and most likely took a few years to happen. For example, we do not observe any change in the average level of album fit between 2001 and 2005 compared with the preshock period. Similarly, we do not see any systematic change in the average sonic features of songs within each genre. Finally, we do not see any sorting by artist quality into experimentation with fit postshock.

That said, our evidence, particularly for the relationship between misfit and appearance on the billboard chat, is correlational. Therefore, we caution against interpreting the results of our archival analysis as causal. Our laboratory experiment was partly motivated to alleviate these causality concerns.

Control Variables. We use several control variables in our analysis. First, to ensure that our results are not driven by outlier songs, that is, by songs that do not fit other songs on their corresponding album, we control for the average cosine similarity between the focal track and other tracks on the album. We also control for length of each track, the track's position on the album, number of tracks on the album, and the primary genres assigned to the album. We also control for how strongly a track represents its genre (or deviates from it) by including a measure of cosine similarity between the track's vector of sonic features and the vector of average sonic features for its primary genre.[6] We also include both artist and year fixed effects to control for the idiosyncratic characteristics of artists and the general macro trends in music production and consumption.

Estimation Model for the Archival Analysis

To test our first prediction, we use the following model to estimate the effect of fit between tracks on track performance:

$$Billboard\_Appearance_{iat} = \beta_0 + \beta_1 Album\_Fit_i + X_{it} + A_a + T_t + e_{iat},$$

where i indicates songs, a indicates artists, t indicates year, A_a indicates artist fixed effects, T_t indicates year fixed effects, and X_{it} is the set of aforementioned control variables.

Next, to examine how the results change after the shift away from album consumption in 2001, we use the following model:

$$Billboard\_Appearance_{iat} = \beta_0 + \beta_1 Album\_Fit_i + \beta_2 Album\_Fit_i \times Post\_2001_t + X_{it} + A_a + T_t + e_{iat},$$

where $Post\_2001_t$ is a dummy equal to one for years after 2001 and zero otherwise. We cannot include *Post_2001* as a separate variable in the regression because its effect will be fully absorbed by year fixed effects. Main results are produced using linear ordinary least squares (OLS) estimations. Given that the fit measure is constructed at the album level, standard errors are also clustered at the album level.[7]

We further use an OLS regression to estimate the effect of album fit on the number of weeks that a song appears on the chart, including the same set of independent and control variables as before.

Archival Analysis Results

Table 1 reports the descriptive and summary statistics for the variables in our analysis. The typical album in our sample has about 13 tracks. The average track lasts about four minutes and has about 0.5% chance of appearing on the Billboard Hot 100 chart. The average album has a high level of fit between its tracks, at 0.84 cosine similarity score, suggesting that album tracks are often quite similar to each other in terms of their sonic features. The average track also closely resembles the prototypical features of its genre. The average cosine similarity between tracks' sonic features and the average sonic features of their genre is 0.905. Table A1 in the Online Appendix reports the correlations between all variables.

Column 1 in Table 2 reports the association between album fit and the likelihood of each song in the album appearing on the Billboard Hot 100 chart. Consistent with our first prediction, the estimated coefficient of *Album fit* suggests that a one-standard-deviation increase in fit between tracks on an album increases the odds of each song appearing on the Billboard Hot 100 chart by approximately 16%. In column 2, we introduce the interaction with the *Post_2001* dummy. The interaction effect in column 2 suggests that the positive effect of album fit drops significantly after 2001, in line with our section prediction. The estimated effect suggests that while a one-standard-deviation increase in fit between tracks is associated with a 24% increase in the chance of each song appearing on the Billboard Hot 100 before 2001, this effect drops to only 9% after 2001, a decline of 15 percentage points relative to the pre-2001 odds.

Columns 3 and 4 report the results for the second dependent variable, the number of weeks that a track stays on the Billboard Hot 100. Again, the results in column 3 show a positive association between the level of fit at the album level with the likelihood that each song in the album would appear on the Billboard Hot 100 charts. A one-standard-deviation increase in album fit is associated with an increase of 16% in the number of weeks a song stays on the chart. The estimates in column 4 further suggest that, although a one-standard-deviation increase in album fit increases the number of weeks that a song appears on the chart by approximately 24% relative to the baseline before 2001, after 2001 this relationship drops by 16 percentage points, to roughly 8%.

The results confirm our theoretical prediction that in contexts where an individual element is consumed as part of a whole, the extent of fit between the composing elements affects how consumers evaluate the quality of the focal element. Moreover, the findings are consistent with the idea that as the salience of the whole, that is,

Table 1. Descriptive Statistics (Archival Analysis)

Variable	Observations	Mean	Standard deviation	Minimum	Maximum
Album level					
Number of tracks	31,925	12.802	2.871	8	20
Album fit	31,925	0.840	0.061	0.425	0.993
Track level					
Appearance on Billboard Hot 100	359,040	0.006	0.076	0	1
Number of weeks on Billboard	359,040	0.048	0.645	0	9
Track similarity	359,040	0.843	0.089	0.087	1.000
Track length (minutes)	359,040	3.969	1.320	1.000	9.200
Order of track on the album	359,040	7.228	4.227	1	20
Similarity to genre	359,040	0.905	0.058	0.330	0.997

Table 2. Archival Analysis Results

	(1) OLS *Appearance on Billboard Hot 100*	(2) OLS *Appearance on Billboard Hot 100*	(3) OLS *No. of weeks on Billboard Hot 100*	(4) OLS *No. of weeks on Billboard Hot 100*
Album fit	0.016***	0.024***	0.123**	0.191***
	(0.006)	(0.007)	(0.050)	(0.059)
Album_Fit × Post_2001		−0.015**		−0.124**
		(0.007)		(0.061)
Track similarity	−0.003	−0.003	−0.017	−0.017
	(0.002)	(0.002)	(0.019)	(0.019)
Track length (minutes)	0.000***	0.000***	0.003***	0.004***
	(0.000)	(0.000)	(0.001)	(0.001)
Number of tracks	0.000***	0.000***	0.002***	0.002***
	(0.000)	(0.000)	(0.001)	(0.001)
Track order	−0.001***	−0.001***	−0.005***	−0.005***
	(0.000)	(0.000)	(0.000)	(0.000)
Similarity to genre	0.002	0.002	0.013	0.012
	(0.003)	(0.003)	(0.024)	(0.024)
Constant	−0.005	−0.012*	−0.037	−0.094*
	(0.005)	(0.006)	(0.046)	(0.053)
Genre fixed effects	Yes	Yes	Yes	Yes
Artist fixed effects	Yes	Yes	Yes	Yes
Year fixed effects	Yes	Yes	Yes	Yes
Number of observations	359,040	359,040	359,040	359,040
Adjusted R^2	0.018	0.018	0.018	0.018

Notes. The analysis is at the track level. The estimation is based on OLS regression with robust standard errors clustered at the album level.
***$p < 0.01$; **$p < 0.05$; *$p < 0.1$.

albums in this context, drops, the misfit bias would also fall.

We ran several additional robustness checks. Figure A1 in the Online Appendix shows the yearly effects of album fit on the likelihood of a song appearing on the Billboard Hot 100, confirming that the drop in the effect of album fitness begins in 2001 and continues until 2005. Table A2 further shows the analysis with various alternative estimation models including logit with conditional fixed effects, OLS with errors clustered at the artist level, and OLS with errors clustered at both artist and album fixed effects. The results are again in line with those reported in Table 2.

We also conducted analyses to address various alternative explanations that may explain the drop in the impact of fit after 2001. First, one might be concerned that artists responded to the 2001 shock by changing the level of fit among their songs on an album. Figures A2 and A3 in the Online Appendix show that the average level of fit between tracks on albums and the sonic features of tracks in our sample do not change in any material way after 2001, suggesting that the results are unlikely to be driven by systematic changes in how artists produced music. There is also the possibility that only higher-quality artists engaged in more experimentation after 2001, essentially reducing the level of fit on their albums. As a result, what we observe as the decline in the impact of fit might alternatively be explained by an increase in the quality of albums that have higher misfit after 2001. To test this idea, we examined how album fit changed differentially for artists who already had at least one hit on the Billboard Hot 100 chart during the pre-2001 period in our sample. Table A3 in the Online Appendix reports the results of this analysis. The interaction between an artist's having a Billboard song during the preshock period and the post-2001 dummy is not significant and is also in the opposite direction of what this alternative explanation would suggest.

Laboratory Experiment

The main results presented in the previous section support our arguments. However, they have two limitations. First, despite various robustness checks, our empirical design does not allow for a strong causal interpretation. To establish true causal evidence, we should compare the same song with itself having higher or lower album fit, something impossible to do in our observational data. Second, the results cannot rule out the possibility that individuals may penalize tracks on a less-fit album based on inferences about the quality of the album producer, such that a song is perceived as worse because the producer releases inconsistent songs. Therefore, the lower success of tracks on albums with poor fit might be driven by lower perceptions of producers' quality rather than fit per se.

To address these issues, we conducted an online laboratory experiment. The experiment further allows us to

triangulate our findings from the archival analysis in a more controlled environment. It also provides additional evidence for our prediction in a different setting from the music industry—art galleries. Note that one reason for choosing photographic galleries for the experiment, instead of music albums, is that a gallery can be consumed much more quickly. Asking experimental participants to listen to randomly constructed albums would not have been economically feasible at scale given the time required, nor would we have been able to ensure participants actually listened to the albums.

In the laboratory experiment, we presented participants with a gallery of three photos and asked them to rate how much they liked the gallery as well as each of the photos in the gallery individually, on a 0–10 scale.[8] The design of our outcome variables is based on the increasingly common consumer ratings found on online platforms (e.g., IMDb, Goodreads, Yelp) that have been studied in prior work (Sharkey et al. 2023). Participants were provided the following instructions at the start of the experiment:

> A few professional photographers have each sent us one photo from their portfolios. We would like to create a gallery composed of three photos from this set of photos. To do so, an algorithm has generated all possible galleries that can be created by randomly picking three photos from the set. We would like to get your opinion on one of these galleries and the photos included in it.

The goal of this prompt was to minimize the potential for participants to draw inferences about unobserved producer quality on two fronts. First, it informs participants that each photo was sent to us by a separate photographer, meaning that the quality of one photo should not be directly related to the existence of another photo, one of the mechanisms already explored in the literature on the perceptual effects of category spanning (Negro and Leung 2013). Second, it highlights that the fit between the three photos is random and that no producer has been involved in actively curating the galleries.[9]

The photos used in the study were acquired from either online photography communities (Flickr.com or FStoppers.com), or iStock, a well-known stock image provider. In the case of the former, we acquired explicit consent from each photographer to use their photos for the purpose of our experiment. In the case of iStock, we purchased the rights to the photos directly from the website. The perceived quality of photos was tested in a pilot study to ensure that none of the photos were outliers.

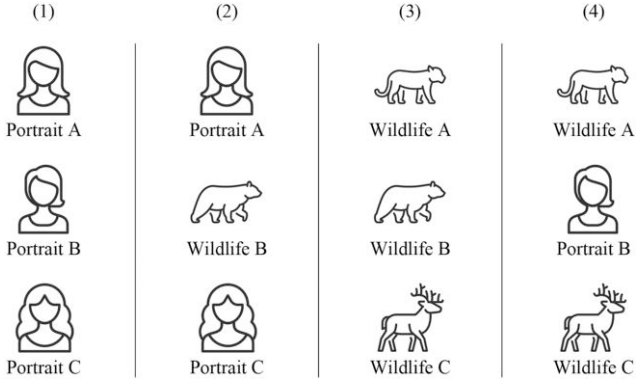
Six photos were used in the final materials. Three were black-and-white portraits of women, and three were color photos of wild animals. The participants were randomly allocated to either a "fit" condition or a "misfit" condition. Participants in the fit conditions either saw only the three black-and-white portraits or only the three wildlife photos. Participants in the misfit conditions saw either two portraits (chosen randomly from the sample of three) and a wildlife photo (also chosen randomly), or two wildlife photos (again chosen randomly) and one portrait (again chosen randomly). The order of photos in all conditions was randomized to control for the potential ordering effect on evaluations. Figure 2 shows four different examples of galleries that participants in our experiment evaluated.

We also asked participants to explicitly evaluate the fit between the photos in the gallery they had evaluated (i.e., "How well do you think these pieces fit together?"). The perceived fit of the misfit galleries was significantly lower than that of the fit galleries, as reported further below.

We are interested in the difference between the average evaluation of each photo conditioned on whether the photo is shown in a fit or misfit gallery. By including photo fixed effects, we can essentially calculate the change in the evaluation of photos based on whether they appear in a fit or misfit gallery.

Moreover, having two sets of photos (portraits and wildlife) allows us to ensure that our results are not driven by direct spillover effects. To illustrate how our experimental design addresses this concern, consider the comparison between a fit gallery composed of three portraits and a misfit gallery of two portraits and a wildlife photo. Example Galleries 1 and 2 of Figure 2 depict two specific examples of galleries reflecting this description. Note that each participant evaluates one only of the galleries, and that other galleries reflecting the description are possible given our randomization of photos and ordering. For our example, assume that the observed average rating of the photo of the bear in Gallery 2 is lower than the rating of the middle portrait in Gallery 1. The negative opinion about the bear in

Figure 2. Four Examples of Galleries Rated by Participants



Note. The original photographs used in the experiment are not reproduced here because of copyright, but the Online Appendix includes links to each source.

Gallery 2 might spill over to the top and bottom photos in that column. As a result, what we might interpret as the negative effect of misfit in Gallery 2 might instead be driven by a direct spillover effect. However, in our experimental design, this issue is addressed by including another set of photos that mirrors the same configuration but instead is composed of three wildlife photos and an outlier from the portrait set to create the fit and misfit galleries. Example Galleries 3 and 4 in Figure 2 illustrate these cases; again, note that the specific order and choice of the outlier photo are completely randomized in our experiment. The spillover mechanism explained above should now lead to a relative positive effect on the top and bottom photos in Gallery 4, because the portrait photo in the middle of Gallery 4 was assumed to be liked more than the bear photo by the average participant. Our theory based on the impact of misfit offers a different prediction, essentially arguing that the evaluations of the top and bottom photos in both Galleries 2 and 4 would be lower than for the same photos in columns 1 and 3, respectively. Therefore, having both sets of photos allows us to pool the evaluations across all comparisons and thus cancel out the potential effects of any direct positive or negative spillovers driven by comparisons of photos within each gallery.

In order to test whether emphasizing the whole can exacerbate the misfit bias, we further randomized the order in which participants were asked to rate the gallery, the individual photos, and the fit. Note that in all cases, participants were provided identical instructions at the start of the experiment and presented with a single gallery. However, in line with our prediction, we expect participants that were asked to evaluate the gallery first to exhibit larger misfit biases when evaluating individual photos. In the same fashion, we expect individuals who were asked to evaluate the fit among the photos before evaluating the photos to also show a greater level of misfit bias, assuming that an evaluation of fit prompts individuals to think more holistically about the gallery.

For this experiment, we recruited 2,500 participants from Prolific, evenly divided between the fit and misfit conditions. Participants were required to reside in a majority English-speaking country and be fluent in English. The majority of participants were from the United Kingdom or the United States. Those who had participated in any pilot study were excluded from the main study. We excluded participants that failed the attention tests or were extreme outliers in terms of time spent on the experiment (using mean plus two standard deviations as the cutoff). The final sample includes 2,330 participants. Basic demographics are reported in Figure A4 of the Online Appendix.

To estimate the impact of misfit on participants' evaluations of photos, we use two different approaches. Given our across-subject experimental design, where each participant views a fit or misfit gallery, in the first approach we calculate the average ratings of photos evaluated by each participant. Our basic finding is based on the comparison of these mean ratings between participants that saw each type of gallery. Given that Prolific provides a number of participant-level control variables, we also regress the average photo ratings on whether the participant was shown a fit or a misfit gallery using the following estimation model:

$$Evaluation_i = \beta_0 + \beta_1 misfit\_gallery_i + X_i + e_i,$$

where $Evaluation_i$ indicates participants' average evaluation of the three photos they were shown, and $misfit\_gallery_i$ is a dummy variable indicating whether the participant was shown a fit or a misfit gallery. We include controls for participants' age, gender, employment status, nationality, and how long they took to complete the experiment, as indicated by X_i. We use the same estimation model to also estimate the impact of misfit on participants' evaluations of the gallery shown to them and the fit among the photos, essentially testing the baseline expectation from prior research and the validity of our fit intervention.

One potential issue with this approach is that, for misfit galleries, the average rating of photos incorporates the rating of the outlier photo in the gallery. Therefore, an observed lower average rating can be driven by participants penalizing the outlier photo rather than penalizing the other two photos because of misfit bias. To address this issue, we further restructured the data to the participant-photo level, essentially creating three observations per participant, each observation corresponding to each of the photos the participant has rated. With this approach, we can directly exclude outlier photos from the analysis and include individual fixed effects for each photo, as well as a control for the order of the photo in the gallery. We use the following estimation model for this analysis:

$$Evaluation_{ip} = \beta_0 + \beta_1 misfit\_gallery_{ip} + X_i + P_p + e_{ip},$$

where $Evaluation_{ip}$ indicates participant i's evaluation of photo p, $misfit\_gallery_{ip}$ indicates whether photo p was shown in a misfit gallery to participant i, X_i is the same set of controls as before for participant i, and P_p indicates the photo fixed effects as well as a control for the photo order in the gallery.

To test our second prediction (whether emphasizing the whole exacerbates the misfit bias), we add an interaction between the misfit indicator and whether the participant evaluated the gallery before the photos and whether the participant evaluated the fit among the photos before the photos. In both cases, we expect the interaction term to be negative, indicating a worsening of misfit bias.

Laboratory Experiment Results

Table 3 reports the raw results and descriptive statistics for the main variables summarized by the fit and misfit gallery conditions. First, galleries with three photos of the same type ("fit galleries") were perceived to have considerably more fit than the galleries that included an outlier photo ("misfit galleries"), confirming the validity of our intervention. Consistent with prior research, misfit galleries were rated lower overall than fit galleries. Our raw main result is reflected in the comparison between the average individual photo ratings in the two types of galleries. Consistent with our predictions in this paper, the individual photos, on average, were rated lower if they appeared in misfit galleries compared with fit galleries; this was also true when excluding the ratings of the "outlier" photos in misfit galleries. The comparison of means tests for all these differences are significant ($p < 0.001$).

We report the regression version of these results in Table 4. Results in column 1 confirm that participants could clearly recognize the lower level of fit among photos in misfit galleries, rating the fit among the photos 3.812 points lower when they were presented in a misfit gallery than a fit gallery.

The estimated effect in column 2 suggests that participants rated misfit galleries 1.324 points lower than misfit galleries, a drop of approximately 18% compared with the baseline evaluation of 7.274 for fit galleries. As expected, the results confirm prior research suggesting that consumers on average penalize products with misfit elements.

Results in column 3 show the impact of misfit on the average rating of photos. Participants on average have rated photos in misfit galleries half a point lower than those in fit galleries, a drop of approximately 7% relative to the baseline average rating of photos in fit galleries.[10] In columns 4 and 5, we introduce the interaction with whether participants evaluated the gallery before the photos and whether participants evaluated the fit before the photos, respectively.

The model in column 4 reports the effect of a participant being asked to rate the gallery before versus after rating the individual photos. Participants who were asked to rate the photos before rating the gallery rated photos in misfit galleries 0.327 points lower than those in fit galleries. Participants who were asked to evaluate the gallery first before the photos, however, evaluated the photos in misfit galleries 0.621 points lower than those in fit galleries, a 9% drop relative to the average ratings of photos in fit galleries.

Similarly, the model in column 5 reports the effect of a participant being asked to rate gallery fit before versus after rating the individual photos. Participants who were asked to rate the photos before evaluating the fit among them rated photos in misfit galleries 0.336 points lower than those in fit galleries. The misfit bias again intensified when participants were asked to evaluate the fit among those in fit galleries (10% lower).

In columns 6–8, we repeat the estimations in columns 3–5, but we use the restructured data at the participant-photo level, excluding the outlier photos and including photo fixed effects, as well as a control for the order of the photo in the gallery. The results are consistent with those reported in columns 3–5. However, the penalty size declines slightly, suggesting that some of the estimated penalty in columns 3–5 were attributed to participants additionally penalizing the outlier elements. The estimated effect of misfit in column 6 suggests that participants rated a fit photo in a misfit gallery 0.342 points lower than the same photo shown in a fit gallery, an approximate 5% decline relative to the baseline of evaluations for fit galleries. The results in column 7 suggests a decline of 0.179 points in evaluation of photos in misfit galleries when participants were asked to evaluate the photos before evaluating the gallery. The penalty increases to 0.452 points when participants were asked to evaluate the gallery before the photos. Similarly, the results in column 8 suggest misfit penalties of 0.200 points and 0.551 points when participants evaluated photos before the fit among them and vice versa, respectively. Overall, the results are in line with those from the archival analysis and support our theoretical predictions.

Discussion

The combination of distant elements is fundamental to the creation of new products, but it is not without cost.

Table 3. Raw Results and Descriptive Statistics (Laboratory Experiment)

Variable	Observations	Mean	Standard deviation	Minimum	Maximum
Fit galleries					
Gallery rating	1,180	7.274	2.171	0	10
Average photo rating	1,180	7.219	1.851	0	10
Fit rating	1,180	7.489	2.192	0	10
Misfit galleries					
Gallery rating	1,150	5.988	2.287	0	10
Average photo rating	1,150	6.765	1.740	0	10
...excluding outlier photo	1,150	6.924	2.200	0	10
Fit rating	1,150	3.691	2.231	0	10

Table 4. Laboratory Experiment Results

	(1) OLS *Fit rating*	(2) OLS *Gallery rating*	(3) OLS *Average photo rating*	(4) OLS *Average photo rating*	(5) OLS *Average photo rating*	(6) OLS *Photo rating*	(7) OLS *Photo rating*	(8) OLS *Photo rating*
Misfit gallery	−3.812***	−1.324***	−0.502***	−0.327***	−0.336***	−0.342***	−0.179*	−0.200**
	(0.092)	(0.093)	(0.075)	(0.118)	(0.096)	(0.061)	(0.096)	(0.079)
Gallery evaluated before photos				0.136			0.144*	
				(0.107)			(0.078)	
Gallery evaluated before photos × Misfit Gallery				−0.294*			−0.272**	
				(0.152)			(0.124)	
Fit evaluated before photos					0.249**			0.240***
					(0.106)			(0.078)
Fit evaluated before photos × Misfit Gallery					−0.415***			−0.351***
					(0.151)			(0.124)
Constant	7.385***	7.010***	7.020***	6.932***	6.908***	7.033***	6.940***	6.927***
	(0.136)	(0.138)	(0.111)	(0.129)	(0.120)	(0.087)	(0.100)	(0.093)
Participant-level controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Photo fixed effects	No	No	No	No	No	Yes	Yes	Yes
Photo order control	No	No	No	No	No	Yes	Yes	Yes
Number of observations	2,330	2,330	2,330	2,330	2,330	5,921	5,921	5,921
Adjusted R^2	0.453	0.112	0.064	0.064	0.066	0.120	0.120	0.121

Notes. The difference between the number of observations between columns 1–5 and columns 6–8 is because of the difference in the unit of analysis. The unit of analysis is at the participant level in columns 1–5. The unit of analysis is at the participant-photo level in columns 6–8. Moreover, outlier photos are excluded from analysis in columns 6–8.

***$p < 0.01$; **$p < 0.05$; *$p < 0.1$.

This paper focuses on one such potential cost, the impact of the combination for the evaluation of the elements combined. We predicted that penalties stemming from distant combinations—misfit—can spill over from the whole to the elements in that whole. We first tested this argument using observational data from the music industry, where, during a very short period, the dominant consumption mode shifted from albums to single songs. We found that individual songs were more likely, on average, to become independent hits when the albums on which they were released presented high levels of fit. In line with our theory, the importance of fit dropped significantly after people started listening to songs independent of their albums. We then tested the argument with an online experiment using galleries of photographs. This allowed us to control for evaluator beliefs about producer quality and to fix the quality of the individual elements being evaluated. Our results show that the same photographs were evaluated less favorably when included in galleries with lower fit.

Limitations

Although our study provides important insights into how misfit at the product level affects evaluations of individual elements, it is not without limitations. One key limitation lies in the scope of our theoretical framework. Although we draw on the halo effect to explain the spillover of negative perceptions of fit onto individual elements, our analysis cannot fully disentangle the specific cognitive mechanisms driving this bias. For example, it remains unclear whether evaluators penalize elements because of an overarching negative impression of the product or because they cognitively conflate fit with the intrinsic quality of the elements.

Our empirical context, focusing on cultural products such as music albums and photographic galleries, also presents a limitation. These settings provide clear measures for both fit and individual element evaluations, but they may not generalize directly to other domains. For instance, in industries where quality metrics are more objective or standardized—such as pharmaceuticals or engineering—the role of fit might differ significantly. Moreover, we use similarity in elements as a measure of product-level fit: in the music study, similarity in acoustic features; in the online experiment, similarity in photo subject matter and style. However, in another context, contrast between elements might be a source of fit. At this moment, research on sources of fit or misfit across contexts is scant. Future work might expand the definition of fit to include more complex relationships between elements, where their degree of complementarity can be explicitly theorized and measured.

Another limitation relates to the production side of misfit combinations. Although our studies seek to isolate the causal effects of misfit on element evaluations, they do not address the strategic choices underlying the creation of misfit products. Little is known about who tends to make such a choice, and under what circumstances. For instance, do producers with higher tolerance for risk or those catering to niche audiences deliberately embrace misfit as a strategy? A deeper investigation of fit and misfit as a deliberate choice

would not only enrich our understanding of the antecedents of misfit but also illuminate the broader trade-offs inherent in creative and innovative work.

Finally, part of our archival analysis is correlational, and despite efforts to mitigate endogeneity concerns, we cannot establish causal relationships with certainty. For example, producers may allocate higher-quality elements to albums with greater fit, creating spurious correlations. Although our experimental design addresses some of these issues, it is conducted in a controlled setting that does not fully replicate the complexities of real-world consumption and evaluation. Future work could explore longitudinal or quasi-experimental designs to strengthen causal inference and examine how these dynamics evolve over time and across markets. By addressing these limitations, future research could build on our findings to refine theories of innovation, creativity, and consumer judgment.

Contributions

Our paper makes several contributions. The findings naturally contribute to the conversations about the potential pitfalls of boundary spanning, expanding on findings related to the perceptual discounts that audiences often exert on producers that span categories (Zuckerman 1999, Hsu 2006, Negro and Leung 2013, Leung and Sharkey 2014). A frequent theoretical explanation in this literature is that consumers draw negative inferences about producers' abilities based on their category spanning. Our findings suggest that a similar perceptual discount can occur at the level of the elements being combined. In other words, we show that individuals' dislike of a product itself is enough to generate a discount that spills over to the product's elements. Theoretically, we highlight the distinct role of elements and that of their fit in explaining consumers' perceptions of a product. We also believe that this is one of the few studies to consider the consequences for elements at the product level despite how often multielement products are observed in markets (courses, albums, galleries, etc.). Note that the current managerial "solution" to the perceptual discount in the literature is either to obscure the identity of the producer (Negro and Leung 2013), for example, by producing under multiple brand names, or to make explicit category labels less salient in the market (Leung and Sharkey 2014). The perceptual discount we identify in this study presents no easy solution given that it occurs at the product level. Indeed, a gallery curator cannot simply display a single piece of art; most university courses cannot be taught in a single session; and a music producer of unfit albums had little recourse before the industry shifted to singles as an economically viable production mode.

Moreover, our focus on elements and how they fit together naturally ties to the innovation literature. For decades, this literature has emphasized the crucial role of recombination for the development of new ideas, new products, and new services (Gilfillan 1935, Schumpeter 1942). Bringing together unrelated, distant elements is a difficult process, often leading to failure, but is also a key driver of breakthroughs (Fleming 2001, Leahey et al. 2017). Hence, studies and popular books have highlighted the benefits of measures, processes, and policies that enable the recombination of otherwise disconnected elements (Saxenian 1994, Hargadon and Sutton 1997, Johansson 2004, Vakili and Zhang 2018, Bikard et al. 2019). This paper highlights a hitherto unrecognized cost of distant combinations. In these combinations, elements might appear to have lower value than if they were presented in more familiar combinations. Put differently, distant combination might not only lead to the failure of the whole; it might also lead to the failure of its parts. By ignoring this cost, researchers and practitioners risk having an overly optimistic view of distant combinations.

Beyond these contributions, we believe our theory might apply to a broader set of organizational contexts beyond the case of product evaluation. More specifically, the logic of a misfit bias applies to any setting where three conditions are met: (1) individual elements are brought together into a larger combination, (2) the performance of the whole depends on the quality of individual elements and the fit between them, and (3) the elements themselves face subjective scrutiny. Various other types of product bundles in noncreative industries clearly meet these requirements. For example, organizations often offer bundles of their products and services to customers. This dynamic is also present in at least two broader organizational problems: (1) the challenge of assessing how individual firm resources or routines contribute to overall firm performance and (2) the challenge of assessing how individual team members contribute to overall team performance. In both cases, the performance of the "whole" may be unambiguous—for example, the amount of firm profit or the total team output—and a function of the quality of each element, as well as the fit between elements. Yet outsiders, and even insiders, are often unable to fully understand the specific relationship between each element (e.g., a resource, or team member) to overall performance. As a result, the individual elements might be at risk for the misfit bias.

In the strategy literature, causal ambiguity is a key mechanism to explain how some firms sustain competitive advantage over time (Dierickx and Cool 1989, Barney 1991). Firms are bundles of assets, but it is often unclear how specific resources or routines contribute to overall organizational performance. Our theory suggests that evaluators are likely to undervalue specific firm resources when they are observed in misfit combinations. Similarly, allocation of credit is shown to be notoriously difficult in teams (Vakili et al. 2022). This has been described as the performance nonseparability problem (Alchian and Demsetz 1972). Uribe et al. (2022)

suggest that managers attempt to evaluate an individual employee's contribution to a team by observing changes in team performance when specific individuals are present versus absent. Our study highlights the importance of such experimentation at the organization level. Indeed, in many settings this type of experimentation is likely a helpful approach to overcoming the effects of the misfit bias and forming more accurate assessments of the quality of individual elements, be that a picture in a gallery, an organizational resource, or an individual team member. Future work might also consider how the magnitude of interdependencies between elements might exacerbate the misfit bias by making it more difficult to understand what is driving fit at the product level.

Acknowledgments

The authors thank audiences at the 2024 Academy of Management and Strategic Management Society annual meetings, the Strategy Unit at Harvard Business School, the 2023 SEI Consortium at Warwick Business School, the Values and Valuation Conference at Harvard Business School, and colleagues at London Business School and INSEAD for helpful feedback on earlier versions of this paper.

Endnotes

[1] What is considered a good or bad fit depends on audience expectations (Bowers 2015). In one setting, individuals may expect combined elements to be highly similar to each other in order to fit. In another setting, they may expect variety and thus consider fit to mean having the right level of diversity. In this paper, we take an agnostic view of what misfit constitutes across settings. Empirically, we rely on data and past research to measure misfit in the two settings we study. Nevertheless, we believe that understanding how individuals form expectations about what constitutes misfit between elements is important for both theory and practice.

[2] There are also mechanical reasons for why combinations with distant or unrelated elements may fail. By definition, creators have less knowledge about the compatibility between distant elements than proximate ones, because such combinations are ex-ante less prevalent in the market (potentially nonexistent). As a result, distant combinations are at a higher risk of failure because of incompatibility between those elements. Nevertheless, our focus in this paper is on the cognitive biases against misfit.

[3] The bias may be further amplified because of a cognitive averaging effect, whereby an initial reaction to the whole serves as the anchor for evaluations of its elements (Troutman and Shanteau 1976, Tversky et al. 1988, Weaver et al. 2012).

[4] Note also that the relative unit sales of albums and singles completely flipped during the industry's transition to digital. For example, in 2000, CD singles represented just 3.5% of CD unit sales, making albums the dominant form. By contrast, in 2006, digital singles represented 95.5% of digital download unit sales, making singles the dominant form (source: riaa.com/u-s-sales-database).

[5] The 11 sonic features used are accousticness, danceability, energy, instrumentalness, key, liveness, mode, speechiness, tempo, time signature, and valence. Askin and Mauskapf (2017, p. 919, table 1) provide definitions of each feature.

[6] The vector of average sonic features for each genre is calculated by taking the average of vectors for all songs that are associated with that genre.

[7] Stata's logit with conditional fixed effects does not allow for clustering standard errors. In our robustness checks, we show that the results are consistent if we use logit with conditional fixed effects clustered at the artist level. We further show that the results are robust to linear estimations with dual clustering of standard errors at both album and artist levels.

[8] The specific wording for the scale was 0 = "Don't like it at all" to 10 = "Like it a lot."

[9] The general concern here is that participants dislike, for example, a portrait photo when it is in a gallery with a wildlife photo because they infer that a photographer cannot produce good portrait photos if they also produce wildlife photos.

[10] Using a Sobel test of mediation, we can confirm that the effect of the misfit (defined as a binary variable at the gallery level) on the evaluation of photos is fully mediated by participants' evaluation of fit.

References

Alchian AA, Demsetz H (1972) Production, information costs, and economic organization. *Amer. Econom. Rev.* 62(5):777–795.

Andrews JC, Netemeyer RG, Burton S (1998) Consumer generalization of nutrient content claims in advertising. *J. Marketing* 62(4):62–75.

Askin N, Mauskapf M (2017) What makes popular culture popular? Product features and optimal differentiation in music. *Amer. Sociol. Rev.* 82 (5):910–944.

Barney J (1991) Firm resources and sustained competitive advantage. *J. Management* 17(1):99–120.

Bikard M, Vakili K, Teodoridis F (2019) When collaboration bridges institutions: The impact of university: Industry collaboration on academic productivity. *Organ. Sci.* 30(2):426–445.

Boudreau KJ, Guinan EC, Karim RL, Riedl C (2016) Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Sci.* 62(10):2765–2783.

Bowers A (2015) Relative comparison and category membership: The case of equity analysts. *Organ. Sci.* 26(2):571–583.

Chai S (2017) Near misses in the breakthrough discovery process. *Organ. Sci.* 28(3):411–428.

Chang S (2023) Two faces of decomposability in organizational search: Evidence from singles versus albums in the music industry 1995–2015. *Strategic Management J.* 44(7):1616–1652.

Covert A (2013) A decade of iTunes singles killed the music industry. *CNNMoney* (April 25), https://money.cnn.com/2013/04/25/technology/itunes-music-decline/index.html.

Cutolo D, Ferriani S (2024) Atypicality: Toward an integrative framework in organizational and market settings. *Acad. Management Ann.* 18(1):157–209.

DiCrescenzo B (2000) Radiohead: Kid A. Retrieved December 5, 2024, https://pitchfork.com/reviews/albums/6656-kid-a/.

Dierickx I, Cool K (1989) Asset stock accumulation and sustainability of competitive advantage. *Management Sci.* 35(12):1504–1511.

Enis E (2020) Everything in its right place: How a perfect 10.0 review of Radiohead's 'Kid A' changed music criticism 20 years ago. *Billboard* (March 26), https://www.billboard.com/music/rock/radiohead-kid-a-pitchfork-review-brent-discrescenzo-2000-9342543/.

Feingold A (1992) Good-looking people are not what we think. *Psych. Bull.* 111(2):304–341.

Fisicaro SA, Lance CE (1990) Implications of three causal models for the measurement of halo error. *Appl. Psych. Measurement* 14(4):419–429.

Fleming L (2001) Recombinant uncertainty in technological search. *Management Sci.* 47(1):117–132.

Fleming L, Mingo S, Chen D (2007) Collaborative brokerage, generative creativity, and creative success. *Admin. Sci. Quart.* 52(3): 443–475.

Fricke D (2000) Kid A. *Rolling Stone* (October 12), https://www.rollingstone.com/music/music-album-reviews/kid-a-185607/.

Gilfillan SC (1935) *Inventing the Ship* (Follett, Chicago).

Guinness World Records (2015) Biggest-selling studio album by a female solo artist. Accessed March 12, 2015, https://www.guinnessworldrecords.com/world-records/383776-biggest-selling-studio-album-by-a-female-solo-artist.

Hargadon A, Sutton R (1997) Technology brokering and innovation in a product development firm. *Admin. Sci. Quart.* 42 (4):716–749.

Hsu G (2006) Jacks of all trades and masters of none: Audiences' reactions to spanning genres in feature film production. *Admin. Sci. Quart.* 51(3):420–450.

Hussey A (2021) Shania Twain: Come on over. *Pitchfork* (October 31), https://pitchfork.com/reviews/albums/shania-twain-come-on-over/.

Johansson F (2004) *The Medici Effect: Breakthrough Insights at the Intersection of Ideas, Concepts, and Cultures* (Harvard Business Press, Boston).

Jones C, Maoret M, Massa FG, Svejenova S (2012) Rebels with a cause: Formation, contestation, and expansion of the de novo category 'modern architecture,' 1870–1975. *Organ. Sci.* 23(6):1523–1545.

Kovács B, Hannan MT (2010) The consequences of category spanning depend on contrast. Hsu G, Negro G, Koçak Ö, eds. *Categories in Markets: Origins and Evolution*, vol. 31 (Emerald Group Publishing Limited, Bingley, UK), 175–201.

Lance CE, LaPointe JA, Fisicaro SA (1994a) Tests of three causal models of Halo rater error. *Organ. Behav. Human Decision Processes* 57(1):83–96.

Lance CE, LaPointe JA, Stewart AM (1994b) A test of the context dependency of three causal models of Halo rater error. *J. Appl. Psych.* 79(3):332–340.

Leahey E, Beckman CM, Stanko TL (2017) Prominent but less productive: The impact of interdisciplinarity on scientists' research. *Admin. Sci. Quart.* 62(1):105–139.

Lena JC (2006) Social context and musical content of rap music, 1979–1995. *Soc. Forces* 85(1):479–495.

Leung MD, Sharkey AJ (2014) Out of sight, out of mind? Evidence of perceptual factors in the multiple-category discount. *Organ. Sci.* 25(1):171–184.

Negro G, Leung MD (2013) 'Actual' and perceptual effects of category spanning. *Organ. Sci.* 24(3):684–696.

Negro G, Koçak Ö, Hsu G (2010) Research on categories in the sociology of organizations. Hsu G, Negro G, Koçak Ö, eds. *Categories in Markets: Origins and Evolution*, vol. 31 (Emerald Group Publishing Limited, Bingley, UK), 3–35.

Nisbett RE, Wilson TD (1977) The Halo effect: Evidence for unconscious alteration of judgments. *J. Personality Soc. Psych.* 35(4):250–256.

Rao H, Monin P, Durand R (2005) Border crossing: Bricolage and the erosion of categorical boundaries in French gastronomy. *Amer. Sociol. Rev.* 70(6):968–991.

Roe B, Levy AS, Derby BM (1999) The impact of health claims on consumer search and product evaluation outcomes: Results from FDA experimental data. *J. Public Policy Marketing* 18(1): 89–105.

Rossman G, Schilke O (2014) Close, but no cigar: The bimodal rewards to prize-seeking. *Amer. Sociol. Rev.* 79(1):86–108.

Saxenian AL (1994) *Regional Advantage: Culture and Competition in Silicon Valley and Route 128* (Harvard University Press, Cambridge, MA).

Schumpeter JA (1934) *The Theory of Economic Development* (Transaction Publishers, Piscataway, NJ).

Schumpeter JA (1942) *Capitalism, Socialism and Democracy* (Harper & Brothers, New York).

Sharkey AJ, Kovács B, Hsu G (2023) Expert critics, rankings, and review aggregators: The changing nature of intermediation and the rise of markets with multiple intermediaries. *Acad. Management Ann.* 17(1):1–36.

Thorndike EL (1920) A constant error in psychological ratings. *J. Appl. Psych.* 4(1):25–29.

Troutman CM, Shanteau J (1976) Do consumers evaluate products by adding or averaging attribute information? *J. Consumer Res.* 3:101–106.

Tversky A, Sattath S, Slovic P (1988) Contingent weighting in judgment and choice. *Psych. Rev.* 95(3):371–384.

Uribe J, Carnahan S, Meluso J, Austin-Breneman J (2022) How do managers evaluate individual contributions to team production? A theory and empirical test. *Strategic Management J.* 43(12):2577–2601.

Uzzi B, Mukherjee S, Stringer M, Jones B (2013) Atypical combinations and scientific impact. *Science* 342(6157):468–472.

Vakili K, Zhang L (2018) High on creativity: The impact of social liberalization policies on innovation. *Strategic Management J.* 39(7):1860–1886.

Vakili K, Teodoridis F, Bikard M (2022) Detrimental collaborations in creative work: Evidence from economics. *Organ. Sci.* 33(5): 1741–1755.

Weaver K, Garcia SM, Schwarz N (2012) The presenter's paradox. *J. Consumer Res.* 39(3):445–460.

Wikipedia (2022) Kid A. Accessed November 30, 2022, https://en.wikipedia.org/w/index.php? title=Kid_A&oldid=1142624514.

Zanella P, Cillo P, Verona G (2022) Whatever you want, whatever you like: How incumbents respond to changes in market information regimes. *Strategic Management J.* 43(7):1258–1286.

Zuckerman EW (1999) The categorical imperative: Securities analysts and the illegitimacy discount. *Amer. J. Sociol.* 104(5): 1398–1438.

Bryan K. Stroube is an assistant professor of strategy and entrepreneurship at London Business School. He received his PhD from the University of Maryland's Robert H. Smith School of Business. His research interests include understanding how audiences of various types make quality judgements.

Keyvan Vakili is an associate professor of strategy and entrepreneurship at the London Business School. He received his PhD from the Rotman School of Management at the University of Toronto. His research interests include creativity and innovation, knowledge recombination, and the relationship between collaboration and innovation.

Michaël Bikard is an associate professor of strategy at INSEAD. He received his PhD from the Massachusetts Institute of Technology Sloan School of Management. His research interests include creativity, innovation, and technology strategy.