An Abundance of Katherines*: The Game Theory of Baby Naming

Katy Blumer[1],
Kate Donahue[1],
Katie Fritz[1],
Kate Ivanovich[1],
Katherine Lee[1],
Katie Luo[1],
Cathy Meng[1], and
Katie Van Koevering[1]

[1]Cornell University

1 Introduction

The most important decision in any child's life happens shortly after they are born and is made entirely without their input or approval - their naming. The name given to a child is traditionally kept throughout their life time and has significant impact on their future life path. This momentous decision is made by parents with little education in the game theory inherently present in the highly competitive field of naming.

We attempt to assist these parents with a simple primer into the game theory underpinning the decision of naming. We will introduce the basic set-up of the naming game and formalize the parameters and incentives, wherein parents have some desired properties of the name. We then describe the pitfalls of the most simple interpretations of these models, particularly the dangers of myopic action. Finally, we present experimental results demonstrating the shift in name distributions under this model. These experiments underline the inherent risks in naming a child and highlight how altering various parameters can change the outcomes of naming strategies.

2 Related works

Surprisingly, no one has ever done any research on naming strategies (so long as you conveniently ignore [4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 16, 17, 18, 19, 20, 22, 23, 25] and likely other work).

3 Model

Naming a child is akin to choosing an outfit for the Oscars. It must be unique enough to stand out - no one wants to show up to the Oscars in the same dress - but it must also be similar enough to be recognizable as a name. Lady Gaga's meat dress is fine for an afternoon, but a child named "Meat Dress" would soon become discontented, if not the plaintiff of a lawsuit. Thus, we model name selection based on the desired "uniqueness" of the name.

*With apologies to John Green [11].

3.1 Name frequency and choice model

First, we present our formal model. Assume there is some set of names $\mathcal{A}$. At a time point i, we assume there exists a discrete distribution of popularity over names

$$f_i(a) = \nu$$

such that $f(a) \in [0, 1], \sum_{a \in \mathcal{A}} f_i(a) = 1$. For simplicity, we assume that every name has unique frequency: that is, $f_i(a_j) \neq f_i(a_k)$ for $j \neq k$.

Next, we model parental preferences. It is well-known that parents are always in complete agreement over the name they would prefer to pick for their newborn child. Therefore, we will treat the parents of each child as a unit, and assume that each set of parents $j \in \mathcal{P}$ has some preference over the proportion of the population that would share the same name as their child (we also assume each parental unit has exactly 1 child). For example, $\mu_j = 0.01$ means that parental unit j wants their child to have a name that is shared by 1% of the population. We will use

$$g(\mu) = p$$

to mean that $p \in [0, 1]$ proportion of parents want a name with popularity μ. For example, if $\mu = 0.1$ and $p = 0.2$, then this means that 20% of parents want a name with popularity 10%. This set-up gives us convenient parameters for the model and just enough Greek letters to sound smart enough for publication.

In general, we will assume that parents are *myopic*, with new parents having no concept of time but perfect access to baby name data. We find this a realistic assumption. Mathematically, parents at time step i will pick the name a that currently is closest to their desired frequency μ_j. Given this assumption, then at time step $i+1$, the proportion of babies who have name a is given by the total fraction of parents for whom name a is closest to their desired frequency.

3.2 Satisfiability

If parents are unable to infer the consequences of their actions and act myopically, then it can immediately be seen that some parents will be deeply unhappy with said consequences: for example, if $g(0.1) = 0.2$ (as in the example above), then the 20% of parents who wished that their name has popularity 10%, will end up with a name that is more popular than they anticipated (when $g(\mu) > \mu$). For instance, a parent might anticipate the name "Kate" would be a pleasantly traditional yet unique name with only moderate popularity. They would be wrong [24].

Conversely, if we had $g(0.1) = 0.05$ (or $g(\mu) < \mu$), then parents would end up with a name that is less popular than they anticipated. If $g(\mu) = \mu$, then we say the name with proportion μ is *satisfied*. Ideally, we would like every name to be satisfied, or $g(\mu) = \mu$ for all $\mu \in [0, 1]$. However, that would give us a distribution with total probability > 1, which is the sort of thing that makes statisticians sad. Instead, we see that the entire naming distribution $g(\cdot)$ is satisfied if every name with *nonzero popularity* $\mu > 0$ has exactly μ fraction of the population that desires this name:

$$\begin{cases} g(\mu) = \mu & \mu > 0 \\ g(\mu) = 0 & \text{otherwise} \end{cases}$$

3.3 Stability

Next, we consider an alternative property that we may wish to have: that the distribution of names be *stable*. If an arrangement is stable, this means that given an existing distribution $f_i(a)$ and a parental preference distribution $g(\mu)$, every name's frequency will be exactly the same at the next time step $i+1$:

$$f_{i+1}(a) = f_i(a) \quad \forall a \in \mathcal{A}$$

The simplest way to achieve stability is to merely assume that every parent would prefer to name their child after themselves. That is, every parent wishes their child to have the same "uniqueness" of naming they do, a sort of inheritability of uniqueness - we name this the Dweezil Principle.

3.4 Extremely Reasonable Assumptions

The above model contains several Extremely Reasonable Assumptions (ERAs). The first ERA is the very conservative assumption that there is only one gender, with all children and all names adhering to the same gender. Thus any child may be given any name, so long as it exists in the names list[1]. Another ERA is the Mayfly Parenthood Assumption, in which all parents perish immediately upon naming their child, which makes the math substantially easier.

4 Illustrative example: power law distribution

In this section, we consider the case where both $f(\cdot)$ and $g(\cdot)$ are given by *power law distributions*.

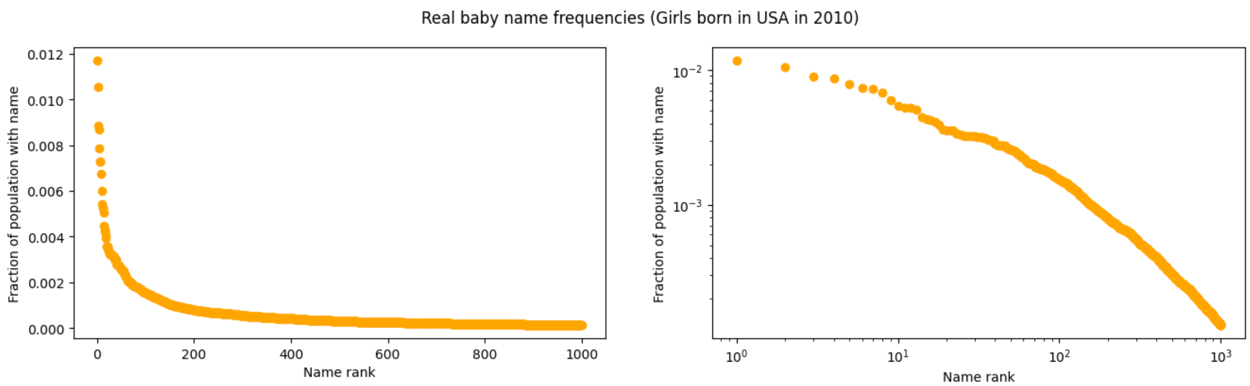

Figure 1: Name frequencies from the Social Security Administration, for girls born in 2010[2]. Note the rough power-law shape.

4.1 Modeling $f(a), g(\mu)$

We begin by defining our variable for name popularity. The popularity of names has been shown to follow a power law distribution [18] (also see Figure 1) . We can model this as:

$$f(a) = K \cdot a^{-t}$$

where a denotes the rAnk of the name within $a \in [1, N]$, t is another constanT, and K is a normalization konstant.

Next, we consider how parents pick the uniqueness of names for their children: the function $g(\cdot)$. Because we are in the Power Law subsection, we will also assume that this distribution of parent preferences is a power law. To minimize our use of variables, we will model this as:

$$g(a) = K' \cdot (a')^{-t'}$$

where a' is the desired frequency of the name (within the range $[\epsilon, 1]$, for $\epsilon > 0$), t' is another constant, and K' is a normalization constant.

[1] If a fixed names list is good enough for the Scandinavians, it's good enough for us [1]

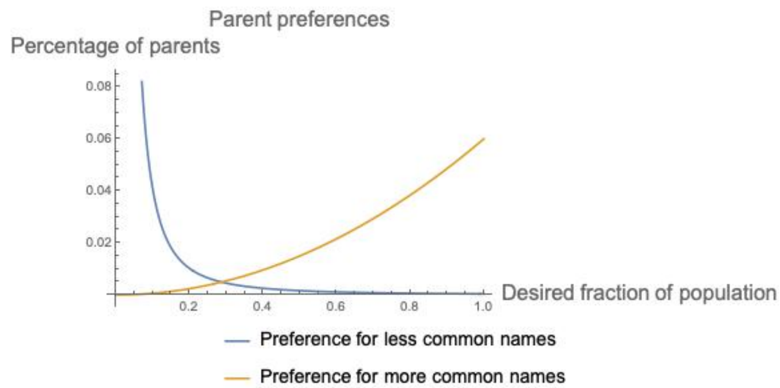

Figure 2: Examples of different parental preferences: a preference for less common names (blue) and preference for more common names (orange).

Note that for $t' > 0$, we have that $g(\mu)$ is decreasing in μ: that is, more parents prefer names that are *uncommon*. Conversely, for $t' < 0$, we have that $g(\mu)$ is increasing in μ: parents prefer names that are *common*.

4.2 Picking names: stability

Given a $f_i(a)$ and $g(\mu)$, the distribution $f_{i+1}(a)$ at the next time step is given by:

$$f_{i+1}(a) = g(f_i(a)) = K' \cdot \left(K \cdot a^{-t}\right)^{-t'} = K' \cdot K^{-t'} \cdot a^{t \cdot t'}$$

Note that this is again a power law distribution[2], with parameter $t \cdot t'$. Next, we can analyze the properties of the resulting distribution.

4.2.1 Picking an uncommon name: a futile quest

First, we consider the case where t' is positive, which corresponds to the case where most parents prefer uncommon names. In this setting, we know that $t \cdot t'$ is also positive, which means that $f_{i+1}(a)$ is *increasing* in a. This tells us that a name with high popularity at time i will have *low* popularity at time $i+1$. Given an original distribution of frequency over names shown in blue in Figure 4, if parents have a preference for less common names, the resulting distribution will look like the orange curve: the least popular name has suddenly become extremely popular, and what was popular at time step i has become horribly passé by time step $i+1$. This means that names will see-saw in popularity from one time step to another, which explains why both your grandmother and niece are named Mabel (Figure 3).

4.2.2 Picking a common name: naming event horizon

However, parents might believe there are benefits to their child sharing a name with others (see [24]). We can model this with the case where t' is negative, which means that most parents would prefer names that are relatively popular, with only a few parents preferring names that are less common. Mathematically, this implies that $t \cdot t'$ is also negative, which means that $f_{i+1}(a)$ is *decreasing* in a. This means that popular names at time step i are also popular at time step $i+1$: the relative order of name popularity stays the same (as shown in the green curve in Figure 4).

If this process is repeated n times, with the same parental preferences, then the resulting power law distribution would have exponent $t \cdot t'^n$. For $t' > 1$ (the "event horizon"), the means that the most popular

[2]This is why we like power laws.

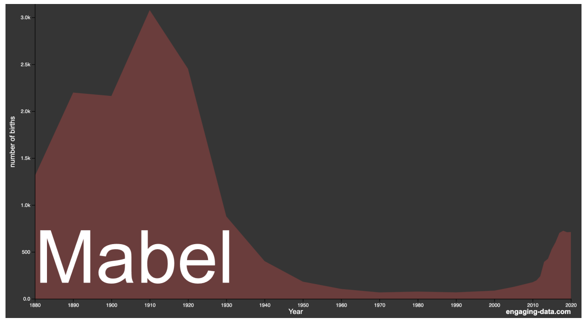

Figure 3: Frequency of name Mabel (image from [3]).

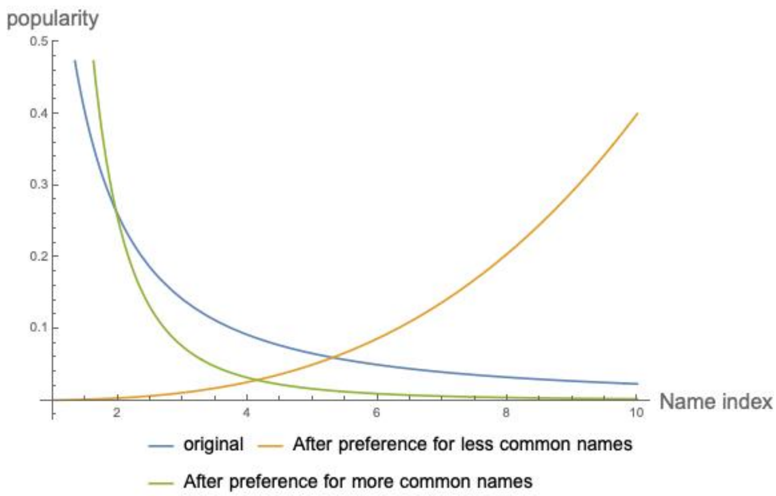

Figure 4: Examples of naming frequency at time step $i + 1$: the original distribution (blue), the distribution after a preference for less common names (orange), and a preference for more common names (green).

names gobble up almost all of the population, resulting in a black hole of names wherein infinite density will eventually fall upon just one name.

5 Simulations

For simulations we use a log-normal distribution of parent preferences, rather than a power law as in Section 4, because a certain author was having issues with SciPy. We baselessly claim a log-normal makes sense because name "uniqueness" is logarithmic; that is, a name belonging to 0.01% of the population is roughly twice as unique as a name belonging to 0.1% of the population (when comparing to a baseline name with 1% popularity).

We work with five different parent preference distributions, with popularity modes of 1% to 0.0001% (Figure 5). From each preference distribution, we generate a sample of 1 million parents, who each choose the name which is currently closest to their desired popularity. The resulting distribution of names is shown in Figure 6. If parents prefer more unique names, the name distribution flattens, since there are many more names at the low-popularity end of the scale for parents to divide between. If parents prefer very popular names, the distribution is even more heavily weighted at the popular end than the original power law distribution. The preference distribution centered at 0.1% in fact closely resembles the original power law distribution.

Measures of parent error (distance from goal popularity) don't show particularly interesting patterns, but are included because they look like dinosaurs (Fig. 7).

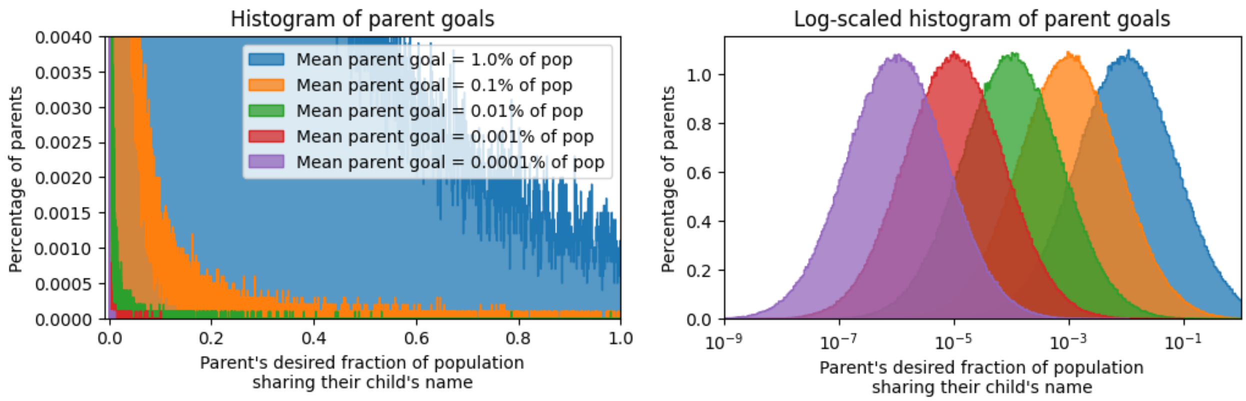

Figure 5: Parent preference distributions used in experiments. Logarithmic x-axis on the right, showing the log-normal distribution shape. Each histogram represents a sample of 1 million parents from the given distribution.

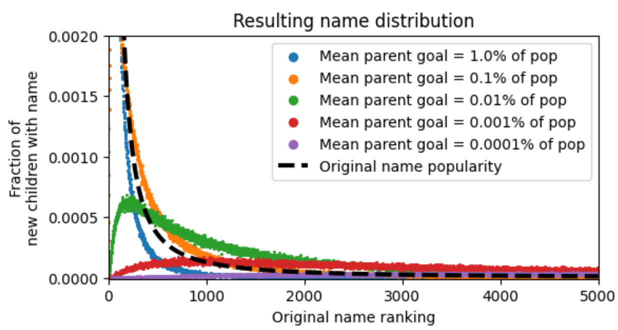

Figure 6: Resulting name distributions for children of parents with given preference distribution.

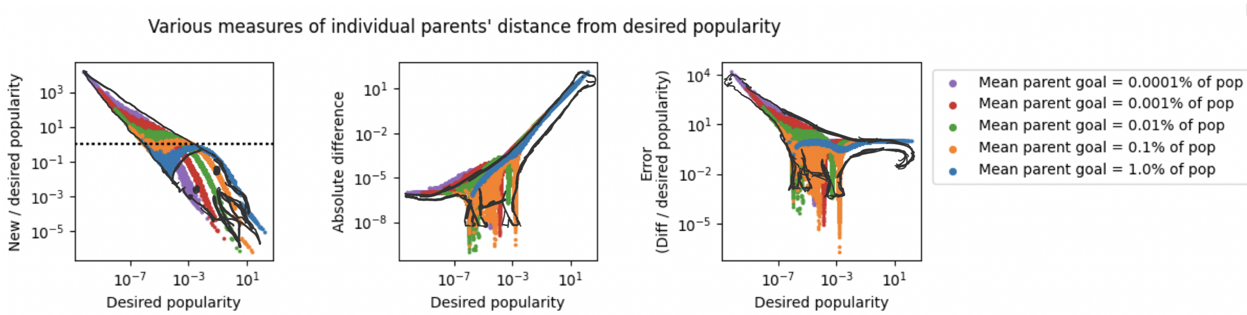

Figure 7: Various measures of parent error, meaning the difference between their chosen name's popularity and their desired name popularity. From left to right: ratio (true popularity divided by desired popularity), absolute difference (absolute value of true−desired popularity), and error (absolute difference divided by desired popularity). Note the adorable dinosaurs.

6 Obligatory Kat-GPT experiment

Because this paper was written in 2024, we include an obligatory section involving generative AI and LLMs. It is fortunate that the most popular LLM of the year appears to be custom-made for this experiment: Kat-GPT. Specifically, we asked Kat-GPT to give us its top ten names for a) a girl, b) a boy, and c) a gender-neutral name (see Figure 8). Then, we calculated the frequency of those names within their gender

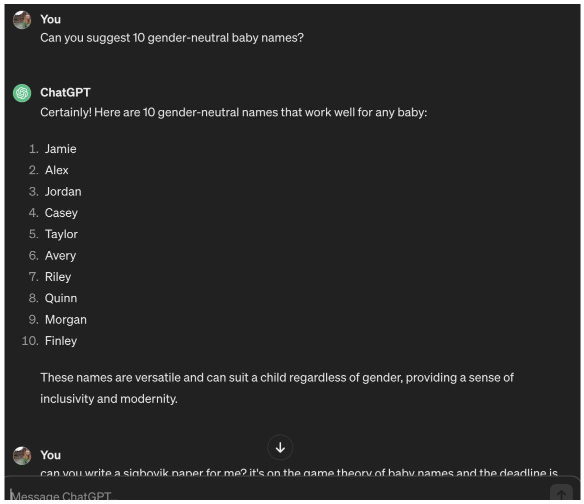

Figure 8: Experiments with Kat-GPT

category (or within the total population, for gender-neutral). The results are given within Table 1.

The difference in mean popularity between "girl" and "gender-neutral" names and between "boy" and "gender-neutral" names are both statistically significant at the $p \leq 0.05$ level, while the difference in mean popularity between "girl" and "boy" names is statistically significant at the $p \geq 0.05$ level.

	"girl" names	"boy" names	"gender-neutral" names
Mean popularity	0.6728	0.666	0.126
Std of popularity	0.134	0.212	0.0619

Table 1: Popularity of names given by ChatGPT ($N = 10^{-e^{\pi \cdot i}}$ queries for each category)

7 Extensions & Future Work

In this section we include multiple extensions that we considered but ultimately were too lazy to actually finish.

7.1 Creation of new names

One of our Extremely Reasonable Assumptions is that there was a fixed list of names. However, new names have occasionally been documented in the wild [21]. One extension could consider a strategy whereby parents could pick a name a' that has some distance $d(a, a')$ from an established name, and derive cost relating both to the popularity of the "base name" a and the distance $d(a, a')$.

$$\min_{a, a' \in \mathcal{A}} \left| K \cdot a'^{-t} - \mu_i \right| + \lambda \cdot d(a, a')$$

where $d(a, a')$ is a distance metric between names (e.g., see Figure 9).

7.2 Non-myopic parents

Another one of our Extremely Reasonable Assumptions was that parents are myopic: that is, if they have a desired name frequency a', they simply blindly pick the name that currently has frequency a', which can

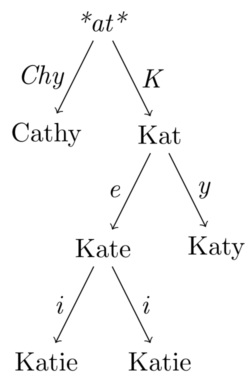

Figure 9: An example of name mutation, where multiple novel (but related) names are displayed with their edit distance from each other.

lead to over- or under-shooting their desired frequency. However, emerging research [15] suggests that people in fact reason strategically about their actions; if this is borne out, considering non-myopic parents may be an interesting avenue of future research (i.e. not us).

8 Conclusions and implications

The science of naming has a long and illustrious history that we didn't bother to look at. Instead, we arbitrarily assigned a new(?) model to describe how parents ought to name their children - namely probabilistically. This model has interesting implications, most interestingly that all naming strategies are futile We also print some plots, for both educational and entertainment purposes, which further emphasize these points and have some nice dinosaurs. But overall, we find only one rule really matters when naming a child: when in doubt, name it Kat[i?e|y].

9 Acknowledgements

We thank F.D.C. Willard for helpful discussions. Simon Shindler contributed significantly to the aesthetic of Figure 7, but could not be named an author for obvious reasons.

We also want to thank the many dozens of people who have confused us for one another at conferences. We enjoyed meeting you.

Finally, we thank Katie's mom for editing this paper and the rest of the parents for making the provably optimal (see above) name choice.

References

[1] Some countries have a list of preapproved baby names. *Interesting Facts*.

[2] Popular baby names, 2024. https://www.ssa.gov/OACT/babynames/index.html.

[3] US baby name popularity visualizer, 2024. https://engaging-data.com/baby-name-visualizer/.

[4] Acerbi, A., and Bentley, R. A. Biases in cultural transmission shape the turnover of popular traits. *Evolution and Human Behavior 35*, 3 (2014), 228–236.

[5] Bentley, R. A., and Ormerod, P. Accelerated innovation and increased spatial diversity of us popular culture. *Advances in Complex Systems 15*, 01n02 (2012), 1150011.

[6] Bush, S. J., Powell-Smith, A., and Freeman, T. C. Network analysis of the social and demographic influences on name choice within the uk (1838-2016). *PLoS One 13*, 10 (2018), e0205759.

[7] Cila, J., and Lalonde, R. N. What's in a name? motivations for baby-naming in multicultural contexts. *Contemporary Language Motivation Theory: 60 Years Since Gardner and Lambert (1959) 3* (2019), 130.

[8] Coulmont, B., Supervie, V., and Breban, R. The diffusion dynamics of choice: From durable goods markets to fashion first names. *Complexity 21*, S1 (2016), 362–369.

[9] Glynn, M. A., and Abzug, R. Institutionalizing identity: Symbolic isomorphism and organizational names. *Academy of Management journal 45*, 1 (2002), 267–280.

[10] Golman, R., Bugbee, E. H., Jain, A., and Saraf, S. Hipsters and the cool: A game theoretic analysis of identity expression, trends, and fads. *Psychological review 129*, 1 (2022), 4.

[11] Green, J. *An Abundance of Katherines*. Penguin Young Readers Group, 2006.

[12] Hahn, M. W., and Bentley, R. A. Drift as a mechanism for cultural change: an example from baby names. *Proceedings of the Royal Society of London. Series B: Biological Sciences 270*, suppl_1 (2003), S120–S123.

[13] Han-Wu-Shuang, B., Hua-Jian, C., and Yi-Ming, J. Understanding the rise of unique names: The emphasis on uniqueness matters.

[14] Kessler, D. A., Maruvka, Y. E., Ouren, J., and Shnerb, N. M. You name it–how memory and delay govern first name dynamics. *PloS one 7*, 6 (2012), e38790.

[15] Lasker, E. *Lasker's Manual of Chess: With 308 Diagrams*. Printing-Craft, 1932.

[16] Lee, M. J., Do Yi, S., Kim, B. J., and Baek, S. K. Theory of fads: Traveling-wave solution of evolutionary dynamics in a one-dimensional trait space. *Physical Review E 91*, 1 (2015), 012815.

[17] Leonardelli, G. J., Pickett, C. L., and Brewer, M. B. Optimal distinctiveness theory: A framework for social identity, social cognition, and intergroup relations. In *Advances in experimental social psychology*, vol. 43. Elsevier, 2010, pp. 63–113.

[18] Li, W. Analyses of baby name popularity distribution in us for the last 131 years. *Complexity 18*, 1 (2012), 44–50.

[19] Müller, J. *Analyzing given names: exploring potentials for a personalized name discovery on Nameling. net*. PhD thesis, University of Kassel, Germany, 2020.

[20] Newberry, M. G., and Plotkin, J. B. Measuring frequency-dependent selection in culture. *Nature Human Behaviour 6*, 8 (2022), 1048–1055.

[21] Okasha, E. *Women's names in Old English*. Routledge, 2016.

[22] Stoyneva, I., and Vracheva, V. Demystifying entrepreneurial name choice: insights from the us biotech industry. *New England Journal of Entrepreneurship 25*, 2 (2022), 121–143.

[23] Twenge, J. M., Dawson, L., and Campbell, W. K. Still standing out: Children's names in the united states during the great recession and correlations with economic indicators. *Journal of Applied Social Psychology 46*, 11 (2016), 663–670.

[24] Van Koevering, K., Meng, K., Luo, K., Ivanovich, K., Fritz, K., Donahue, K., Blumer, K., and Lee, K. An abundance of katherines: The game theory of baby naming. *SIGBOVIK* (2024).

[25] Varnum, M. E., and Kitayama, S. What's in a name? popular names are less common on frontiers. *Psychological science 22*, 2 (2011), 176–183.