Science Advances
AAAS

advances.sciencemag.org/cgi/content/full/7/9/eabd4995/DC1

Supplementary Materials for

People accurately predict the transition probabilities between actions

Mark A. Thornton* and Diana I. Tamir

*Corresponding author. Email: mark.a.thornton@dartmouth.edu

Published 26 February 2021, *Sci. Adv.* **7**, eabd4995 (2021)
DOI: 10.1126/sciadv.abd4995

This PDF file includes:

Supplementary Materials

Comparing IMSDb and WikiHow

Although movie scripts are relatively naturalistic stimuli, they are not slices of real life. Science-fiction and fantasy movies often feature physically impossible events. Even more mundane movies still present curated views of reality that routinely omit actions – such as using the bathroom – that do not contribute to the narrative. As a result, the action transitions in estimated from IMSDb may not perfectly reflect natural action transitions.

To assess this possibility empirically, we conducted a post hoc analysis to compare the IMSDb and WikiHow ground truth estimates to each other. These are logical datasets to compare for two reasons. First, both estimate action transitions in similar ways: using text analysis with verbs as proxies for actions. Second, WikiHow provides instructions for how to accomplish real-life goals. As a result, these action transitions are not likely to be fantastical – since there is not much point in providing instructions regarding how to fix a warp drive that has not been invented (yet). WikiHow is also unlikely to omit important but narratively uninteresting steps, since doing so could lead to the failures to achieve the specified goal. These features make WikiHow a more grounded datasets against which to compare the IMSDb ground truth. If these ground truth estimates were uncorrelated, it might call into question to validity of the movie scripts.

Considering the full set of actions measured in each ground truth dataset (not limited to those rated by participants) we observed an overlap of 88 actions. To measure the similarity between the ground truth estimates, we correlated the off-diagonal elements of the log odds transition matrices comprised from these 88 actions. We assessed the statistical significance of this correlation using the Mantel test. We observed a significant correlation of $\rho = .22$ ($p <$

.0001). This indicates a modest but reliable association between action transitions measured in IMSDb movies and normative action transitions measured in the WikiHow instruction sets.

The modest size of this association may be due to the lack of verisimilitude in the movies or due to the normative, rather than descriptive, nature of the instruction sets on WikiHow. It could also stem from less substantive issues such as differences produced by the structure of the datasets (e.g., script lines vs. instruction steps). However, the presence of a significant association indicates that both measures of ground truth tap into at least some overlapping statistical regularities in action transitions.

Action selection in Study 1

Like Study 4, Study 1 selected actions in part based on the association between transition probabilities and ACT-FACT proximity. However, the selection procedure was much weaker and less direct than in Study 4. In Study 1 we selected based on scores from a text analysis conducted in the paper where we derived the ACT-FAST taxonomy (33). These scores were the first step in developing the named, defined, interpretable dimensions of the taxonomy. However, human ratings of the psychological dimensions were only moderately correlated with the raw scores from the text analysis. Since the verbs selected in Study 1 maximized associations with the text scores, the selection procedure only weakly and indirectly constrained them to be associated with the human ratings we later collected.

Moreover, in Study 1 we selected a complete transition probability matrix, in the sense that all elements of the square matrix were considered. In contrast, in Study 4 we selected 100 specific elements within a transition matrix, rather than all possible transitions between the actions considered. There are far more elements than rows/columns of a square matrix – and the choice space for the optimization algorithm grows factorially as a function of the number of

values it chooses among. As a result, the selection procedure had much more latitude to optimize its choices in Study 4. In contrast, the selection procedure in Study 1 was much more difficult to overfit. As a result, we do not believe that the ACT-FAST associations with ground truth are as problematic in Study 1 as in Study 4. However, readers may nonetheless want to interpret these values with a degree of caution.

Study 5'

Study 5' resulted from an unintentional deviation from the registered plan for Study 5 as a result of a programming bug. This bug resulted in participants rating the transition probabilities between an arbitrary subset of 23/80 actions, rather than the 23 actions we intended to present. We re-ran this study to correct this mistake, yielding the Study 5 results reported in the main text. However, in the interests of transparency, we also report the results of the mistaken data collection here as Study 5'.

A set of 120 participants rated the transition probabilities between actions in Study 5'. One participant was excluded for indicate lack of English fluency, leaving a final sample of 119 (54 Female, 64 Male, 1 Other gender; mean age = 37.8, age range = 20-70). Each participant rated 100 out of the 529 possible transitions between 23 actions. Since the error was detected after collecting this initial sample of participants, no secondary sample was recruited for the purposes of ratings the actions on the ACT-FAST dimensions. As such, these data were analyzed only with respect to the accuracy of participants judgements, and not dimensional mediation.

Item analyses – in all four variants described in the main text – were conducted by averaging transition probability ratings across participants and then correlating them with the ground truth derived from the AVA. The correlation between perceived and ground truth transition probabilities was significant with respect to the full transition probability matrix (ρ =

.27, $p < .0001$), the matrix excluding the diagonal ($\rho = .18$, $p = .0002$), and the symmetric component of the matrix ($\rho = .26$, $p = .0016$), but not the asymmetric component of the matrix ($\rho = .10$, $p = .051$). Participants' accuracy was also statistically significant at the individual level (mean $\rho = .18$, $p = 3.6 \times 10^{-29}$, 95% CI = [.15, .20], $d = 1.38$). Thus, despite the inadvertent selection of a suboptimal subset of actions, the results of Study 5' replicated those reported in Study 5 proper.

Mediation of accuracy by individual dimensions

In addition to testing the mediation of accuracy by the overall ACT-FAST, we also tested whether each of the six dimensions independently mediated accuracy. First, we examined whether each dimensions' proximities were associated with transition probability ratings. To test this, we computed the proximities between each pair of actions based on the dimension ratings provided by separate participants. Proximity was defined as the negative absolute difference between ratings of each action on a given dimension. Proximities were calculated separately for each participant in the dimension ratings studies. The proximities were entered into a multiple regression to predict average rated transition probabilities. We used the symmetric components of the transition similarity matrix, since the proximity metric could only make symmetric predictions. The diagonal was also excluded, as distances from one action to itself are trivially zero. The regressions yielded a set of six regression coefficients for each participant with respect to transition probability ratings. We tested whether these coefficients were significantly greater than zero using one-sample t-tests and percentile bootstrapping, with the latter preferred in cases of qualitative difference between the results. Multiple comparisons across the six ACT-FAST dimensions were controlled using the Holm-Bonferroni procedure.

Second, we likewise tested whether proximities on each dimension were associated with ground truth transition probability estimates. This analysis mirrored the analysis of the rated transition probabilities, except that the dependent variables were the ground truth estimates from each study instead of the rated transition probabilities. This portion of the analysis was not carried out with respect to Study 4. The specific transitions examined in Study 4 were selected based on their association with proximity on the ACT-FAST dimensions (although in a separate half of the data). This effectively guaranteed that dimensional proximity would be correlated with ground truth transition probabilities, making this test uninformative.

The third and final component of the mediation consisted of a partial correlation analysis similar to that applied to the overall ACT-FAST space. In this analysis we examined whether the accuracy correlation decreased when including each dimension's proximities in the model. First, we computed the partial correlation between the rated and ground truth transition probabilities, controlling for all but one of the six dimensions. We then subtracted the full partial correlation away from these all-but-one partial correlations to estimate unique contribution of each dimension to mediating the accuracy association. This procedure was repeated leaving out each dimension in turn. Statistical significance was again tested using percentile bootstrapping, with multiple comparisons controlled via Holm-Bonferroni.

As with the overall mediation analysis, the individual dimension mediation analyses were carried out differently in Study 4, due to its design. There was no dimension rating sample in Study 4 – instead, the study used actions/verbs that had been previously rated in Study 1 or in a previous investigation (33). Each participant's transitions ratings were separately regressed upon the dimensional proximities, with the latter based on ratings averaged across the participants who provided them in previous studies. The partial correlation analysis likewise featured separate

regressions for each participant in the transition rating sample. The version of that analysis which we preregistered only tested whether each individual dimension mediated the rating-ground truth association. However, we realized that this was more lenient than in the other studies – in that it did not control for the other five dimensions to isolate the unique mediational contribution of each dimension. Thus, we deviated from our registered plan, and instead tested whether each dimension mediated accuracy over-and-above the contribution of the other dimensions, as in the other four studies.

The results of these analyses are summarized in Figure S1. They indicate considerably heterogeneity in which specific dimensions are carrying the weight of mediation in different studies. Spiritualism was one of the most successful mediators across the board, correlating uniquely with rated transitions in all but Study 4 and with ground truth transitions in all but Study 3. The food dimension was also a potent mediator, although the pattern of its mediation was at times unexpected: in Studies 2 and 3, the closer two actions were on the food dimension, the less likely the transition between them, in both perception and ground truth. This peculiarity may stem from satiation: Studies 2 and 3 featured the longest timescales (hours) and coarsest actions (activity categories) and as such, reports of a food-related activity often entailed an entire meal. Since one would not expect a person to often eat one meal immediately after another, food-to-food transitions would thus be unlikely.

This variability may result from the large differences in the types of actions and characteristic time scales considered between the different studies. However, at least one dimension mediated accuracy in each of the five studies. This suggests that the different dimensions of the ACT-FAST may be necessary to predict action transitions at different levels of temporal or conceptual resolution. Study 4 – which as specifically tailored to test dimensional

mediation – shows the largest effects across most dimensions. In Studies 2 and 3, the Food dimension mediated accuracy, but did so in an unexpected way: there was a negative associated between proximity on this dimension and both rated and ground truth transition probabilities. That is, the closer two actions were on the Food dimensions, the less likely a transition was between them.

Lists of actions

In Study 1, participants rated the following actions: buy, call, cry, drive, enter, fall, fire, fly, get, grab, hit, jump, knock, know, love, marry, mean, open, pull, reach, run, sell, shut, sit, steal, stop, talk, tell, think, want.

In Study 2, participants rated the following actions: Personal care (sleeping, washing, dressing, grooming, toilet), Household activities (cooking, cleaning, maintenance and repair, personal finance, pet and garden care), Care for & helping household members (caring for children, adults, or elderly), Care for & helping non-household members (caring for children, adults, or elderly who are not part of one's household), Work (primary, side-job, or other income-generating activities, job searching), Education (taking classes, studying, or doing research/homework), Consumer purchases (shopping, researching products), Professional & personal care services (using paid childcare, banking, legal, medical, or real estate services), Household services (using cleaning service, home maintenance service, pet, garden, or vehicle services - not done by self), Government services & civic obligations (using police, fire, or social services, or performing jury duty or voting), Eating and drinking (eating and drinking, including wait time but not cooking), Socializing, relaxing, and leisure (socializing, communicating, attending or hosting social events, relaxing, enjoying entertainment), Sports, exercise, and recreation (participating in sports or exercise, or attending a sporting event), Religious or

spiritual activities (attending religious service, engaging in religious practice), Volunteer activities (administrative, social service, maintenance, cultural, or public health and safety volunteering), Telephone calls (calling friends, family, or others, including internet-mediated video calls), and Traveling (Driving, walking, biking, or riding public transportation for the purpose of travel).

In Study 3, participants rated the following actions: sleeping; grooming (e.g., washing, dressing), health-related self-care; housework; food and drink preparation and clean-up; interior home maintenance, repair, and decoration; exterior home maintenance, repair, and decoration; lawn, garden, and houseplant gardening; animal and pet care; household management (e.g., personal finances); caring for and helping household children; activities related to household children's education; helping household adults; caring for and helping non-household children; helping non-household adults; working; taking classes; doing research or homework for class; shopping; using medical or care services; eating and/or drinking; socializing and communicating; attending or hosting social events; relaxing and leisure; arts and entertainment (other than sports); participating in sports, exercise, and recreation; religious and/or spiritual practices; volunteer administrative or support activities; telephone calls.

In Study 4, participants rated the following transitions:

find → grab	look → use	look → sell	jump → remember
grab → stop	wait → show	treat → hear	start → think
think → end	incorporate → stop	hit → fall	end → sell
keep → meet	remember → try	start → love	cry → talk
work → talk	found → find	pull → put	thank → save
happen → send	need → involve	figure → watch	come → calm

mean → think

calm → believe

keep → break

cry → put

forget → end

fall → work

involve → sit

find → see

seek → keep

cry → wait

hide → kill

send → kill

open → find

hope → save

play → seem

think → lose

grab → reach

mean → come

treat → try

let → happen

send → reach

run → shut

lean → die

work → fall

use → buy

stop → thank

excuse → believe

go → run

find → ask

hear → lean

lose → seek

think → mean

break → grab

want → hit

lock → talk

put → work

think → thank

grab → force

leave → agree

sell → look

locate → agree

locate → put

tell → talk

incorporate → lock

jump → understand

understand → cry

help → lock

hear → calm

seek → listen

reach → keep

drive → ask

love → use

cry → turn

involve → eat

locate → remember

know → excuse

produce → treat

agree → understand

shut → force

meet → ask

forget → think

play → thank

treat → run

thank → hope

understand → forget

excuse → seem

lead → answer

stay → think

recognize → seem

find → need

reach → call

run → believe

listen → figure

come → meet

put → calm

hear → bring

In Study 5, participants rated the following actions: answer phone, bend/bow (at the waist), carry/hold (an object), catch (an object), eat, get up, give/serve (an object) to (a person), grab (a person), lie/sleep, lift/pick up, listen to (a person), open (e.g., a window, a car door), pull (an object), put down, read, sit, smoke, stand, take (an object) from (a person), talk to (e.g., self, a person, a group), walk, watch (a person), watch (e.g., TV).

Dimension definitions

The following definitions/descriptions were used in the dimension rating tasks to elicit ratings of action on the six ACT-FAST dimensions.

Abstraction: "Please rate the action below on the psychological dimension of **abstract/social** versus **concrete/physical**. **Abstract/social** actions tend to be general, not directly observable, and to involve social status (reflecting high or low status, or changing relationships). **Concrete/physical** actions are directly observable, specific, and involve moving physical objects through space, arranging things with respect to others, or altering the parts of something."

Creation: "Please rate the action below on the psychological dimension of **creation** versus **crime**. **Creation** actions tend to involve the production or consumption of music (performing, touring, cheering/booing), the internet (browsing, up/downloading), television (filming, auditioning), writing (rhyming, editing), and other digital or physical media. **Criminal** actions tend to involve stealing from or harming others, and the legal (court, trial) or medical (hospital, surgery) consequences thereof."

Tradition: "Please rate the action below on the psychological dimension of **tradition** versus **innovation**. **Traditional** actions tend to be those which people have performed for years,

decades, or longer, like food preparation, violence and punishment, worship, and ceremony. **Innovative** actions tend to be those which involve technology, machinery, and other modern inventions. These actions would include various types of engineering, computer programming-related activities, and advanced medical procedures."

Food: "Please rate the action below on the psychological dimension of **Food**. **Food** actions are those related to the preparation of food through various types of cooking. N**on-food** actions tend to be related to contexts such as conflict and crime, which often preclude food."

Animacy: "Please rate the action below on the psychological dimension of **animate** versus **mechanical**. **Animate** actions are those which are done by living agents, such as people or animals. **Mechanical** actions are those which can be, or tend to be, done by machines, computer programs, or other automated, artificial processes."

Spiritualism: "Please rate the action below on the psychological dimension of **work** versus **worship**. **Work** actions tend to be related to effort, business, money, and management. **Worship** actions tend to be used in the context of religion, poetry or metaphor, and spirituality."

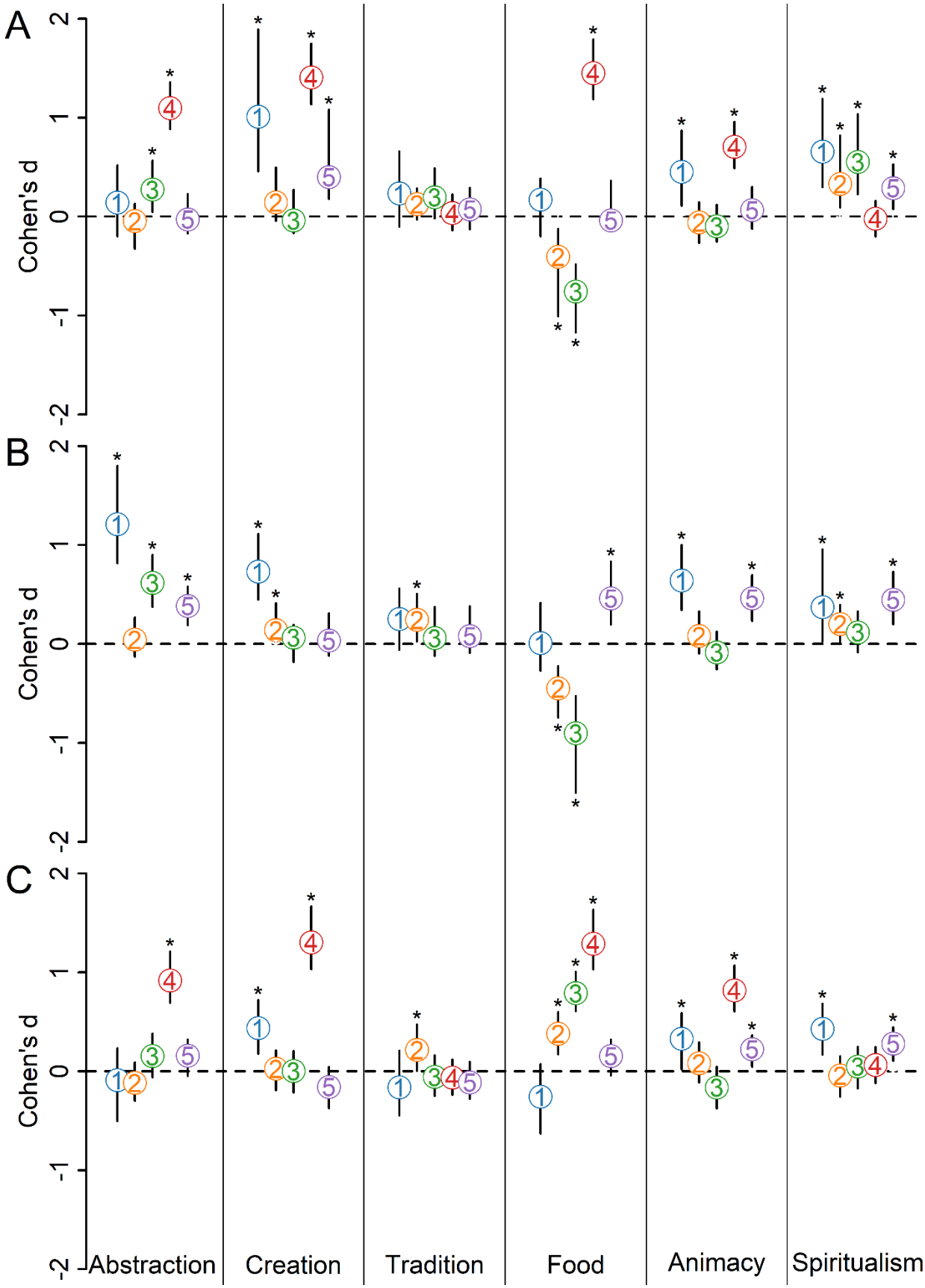
A
B
C
Cohen's d
2
1
0
-1
-2
Abstraction
Creation
Tradition
Food
Animacy
Spiritualism

Figure S1. Mediation of accurate action prediction by ACT-FAST dimensions. Mediation analyses were conducted to determine whether each of the six ACT-FAST dimensions could uniquely explain some portion of the association between rated transition probabilities and ground truth estimates. These analyses consisted of three components: A) associations between ACT-FAST proximities and rated transition probabilities; B) associations between ACT-FAST proximities and ground truth transition probabilities; and C) reductions of rating-ground truth associations by inclusion of ACT-FAST proximities. Numbers and colors indicate Studies 1-5. B) does not feature Study 4, as the association between proximity and ground truth transition probabilities was built in by the selection of actions in that study. This was also the case in Study 1, but the selection was much more indirectly and weakly constrained than in Study 4, and so we tested the associations with ground truth nonetheless. All effect sizes are shown in terms of Cohen's *d* to facilitate comparison across studies, dimensions, and components of the mediation. Error bars indicate 95% bootstrap confidence intervals, Holm-Bonferroni corrected for the six dimensions. Asterisks indicate statistically significant differences from zero.

Table S1. Chance-level accuracy for item analyses.

Study	Full matrix	No diagonal	Symmetric	Asymmetric
1	.090	-.0015	.092	-.0013
2	.058	.0010	.062	-.00070
3	.064	-.00018	.068	.00019
5	.11	-.00039	.11	.0021

Note: The correlations (ρ) expected by chance under the null of no accuracy in each of the item analysis variants. These values reflect the medians of the Mantel test permutation null distributions used to generate p-values for these analyses.

Table S2. ACT-FAST mediates action predictions.

Study	N	Mean Δr	*p*	95% CI	*d*
1	49	.013	.017	[.0033, .024]	.35
2	155	.012	1.1×10^{-6}	[.0073, .016]	.41
3	152	.0095	6.7×10^{-9}	[.0064 .012]	.50
4	156	.21	1.0×10^{-45}	[.19, .23]	1.63
5	153	.0073	.00017	[.0036, .011]	.31

Note: Accuracy – reflected in the correlation between participant-rated transition probabilities and ground truth estimates – was computed separately for each participant using Spearman's ρ. P-values reflect one-sample t-tests on Fisher z-transformed correlations. 95% CI reflect bootstrap percentile confidence intervals on untransformed ρ values. Cohen's *d* values were computed by dividing the mean z-transformed ρ by the standard deviation of the z-transformed ρs.