Novel Goal Creation and Evaluation in Open-Ended Games

Graham Todd*
Computer Science
and Engineering, NYU

Junyi Chu
Department of Psychology
Stanford University

Guy Davidson
Center for Data Science, NYU
and FAIR at Meta

Weijia Xu
Microsoft Research

*Work performed during an internship at Microsoft Research Redmond

Abstract

How do people generate and decide between the wide array of potential goals available to them at any given moment? We study this question in *Minecraft*, a game environment that is both open-ended enough to support a diverse array of goals and structured enough to facilitate quantitative evaluation of different goal features that may impact how people respond to different goals. Specifically, we explore the role of goal familiarity, concreteness, and complexity, which we operationalize using both linguistic analyses and by converting human-generated goals into a programmatic domain-specific language. Our results highlight the unique ways in which game environments like *Minecraft* can facilitate research into how humans engage in open-ended and creative behaviors.

Keywords: goals, modeling, evaluation, games, *Minecraft*

Introduction

Humans invent new goals for themselves across a variety of contexts and timescales, from play to life plans. People can also reason about this vast sea of potential goals under a range of different objectives (e.g., *How fun would this goal be? What would I learn by attempting it?*). Often, people appear to choose goals that are useful, making plans with instrumental or epistemic utility in their current environments (Willatts, 1999; Sim & Xu, 2019). Just as often, however, people select goals with no apparent immediate benefit (Chu et al., 2023), perhaps choosing based on intrinsic notions of fun or curiosity (Schmidhuber, 2010). What information and computations allow people to reason about goals in such flexible ways?

To study flexible goal invention and evaluation, prior work has largely taken one of two approaches. To capture open-endedness, researchers have collected human-generated goals in natural language (Chu et al., 2024). Alternatively, researchers have used virtual game environments that allow goals to be expressed in a programmatic domain-specific description language, which allows analyzing goals in more structured ways (Davidson et al., 2025). Both approaches offer valuable insights but make different tradeoffs: participants may find it more natural to produce and reason about goals expressed in natural language, while virtual game environments can support analysis and interpretation by allowing researchers to impose constraints on possible goals (e.g., by defining available actions and objects).

Presented at Cogsci 2025

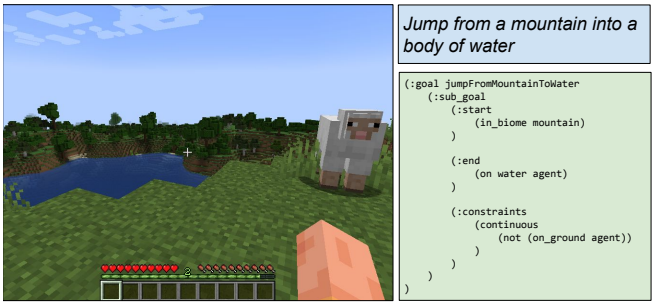

Figure 1: An example frame from *Minecraft*, along with a user-generated goal (in blue) and its translation into our domain specific language (in green).

Here, we aim to get the "best of both worlds" by studying human-generated goals in a video game environment that is open-ended yet structured. We use *Minecraft*, a popular video game in which players navigate and explore a three dimensional environment (Mojang Studios, 2011). Goals in *Minecraft* have the potential to span a wide range of activities, from exploration to combat to social interaction. At the same time, *Minecraft*'s wide playerbase provides a shared grounding for players to understand and evaluate each other's goals as well as computational tools with which to analyze those goals. Our work continues strands of research in cognitive science and artificial intelligence which aim to understand human goals (Dweck, 1992; Elliot & Fryer, 2008; Fishbach & Ferguson, 2007; Molinaro & Collins, 2023) as well as work that uses games (and *Minecraft* more specifically) to study human cognitive capacities (Allen et al., 2024; Shaw, 2023; Lane & Yi, 2017; Canossa et al., 2013; Voiskounsky et al., 2017; Y. Fan et al., 2022).

Here, we investigate both the kinds of goals that people generate and the ways in which they are judged. We begin by exploring how these human-generated goals vary, before examining three kinds of features that may support goal evaluation: (1) *familiarity* (of the goal to the domain and the domain to the participant), (2) *concreteness*, and (3) *complexity*. We treat each question separately, using model-based analyses to determine whether collections of features help explain goal ratings above a baseline. We find support for each hypothesis, with distinct collections of features helping to explain ratings of different goal attributes.

While our results are specific to *Minecraft*, they also help demonstrate the ways in which structured and open-ended game environments can help advance the study of goals more generally.

Environment and Data Collection

Goal Generation Task

Minecraft is a first-person, 3D survival and exploration game developed by Mojang Studios. In it, players are placed into a randomly-generated world composed entirely of different kinds of "blocks" (e.g. stone, wood, iron, water, ...). Players can harvest these blocks and combine them into various useful items using specific "crafting recipes." The game world is also populated with a variety of non-player characters, ranging from the benign (e.g. cows, villagers, squid) to the hostile (e.g. zombies, skeletons, spiders). Notably, *Minecraft* is a "sandbox-style" game – there is no top-down directive to players about their goals or motivations. Rather, players decide for themselves which activities to pursue, bounded only by general mechanics implemented in the game. In the 14 years since *Minecraft*'s release, players have generated and accomplished a vast array of goals, from elaborate recreations of famous landmarks to simulations of *Minecraft* itself built within the game engine.

We recruited $N = 50$ participants using the Prolific platform to propose novel goals that a player might attempt in *Minecraft*.[1] Participants were shown a series of videos and given text explanations of the game's basic mechanics before being asked to produce $k = 5$ goals that (1) a player could reasonably achieve within a few minutes and (2) for which success could be objectively measured. After generating 5 goals, participants were prompted to create a more "complicated, difficult, or interesting" version of each goal, with three examples (for instance, converting the goal "*Drink a potion*" to "*Drink a potion while being chased*"). Finally, we collected basic demographic information from each participant, notably including their recent time spent playing video games in general and *Minecraft* in specific. From this set of 500 goals (250 initial and 250 "complexified") we filtered out 46 which were duplicated other responses or obviously failed to complete the task as well as the goals from 5 participants who did not provide demographic data on their *Minecraft* experience, leaving us with a final dataset of 444 goals (227 initial and 227 "complexified").

Goal Rating Task

To get a better sense of how people evaluate novel goals and the properties of the goals that drive those evaluations, we collect ratings over a subset of the goals generated in Experiment 1. Specifically, we sample 150 of the 444 total goals (75 base goals and their "complexified" versions) by randomly selecting five participants from each of the five self-reported familiarity levels with *Minecraft*. We recruit a separate set of $N = 107$ participants via Prolific and present them with the same information about *Minecraft* as in Experiment 1. Each participant was presented with a random selection of 15 goals (ensuring that each goal received at least 10 independent ratings) and asked to provide ratings along the following axes using a 5-point Likert scale:[2] (1) *How creative is this goal?*, (2) *How interesting is this goal?*, (3) *How difficult is this goal?*, (4) *How fun do you think it would be to attempt to achieve this goal?*, (5) *How much do you think attempting to achieve this goal would teach you about Minecraft?* In addition, participants were asked to provide a brief description of how they would attempt to accomplish the goal and to estimate the number of ways the goal could be accomplished. Our choice of rating criteria largely follows the factors studied in prior work (Chu et al., 2024; Davidson et al., 2025), though we also include the previously-unstudied "teaching" feature.

Programmatic Representation

Prior work has demonstrated that viewing human-created goals as programs in a specialized domain-specific language (DSL) offers a unique and effective lens of analysis (Davidson et al., 2022, 2025). Notably, programs offer a way to standardize the semantic content of a goal against differences in language and evaluate their structure compositionally. To facilitate this kind of analysis, we construct a similar DSL which represents goals in *Minecraft* as temporally extended programs, an example of which is presented in Figure 1. Each goal is represented as one or more sub-goals, which are themselves defined by an ending condition (which must hold for the goal to be satisfied), an optional starting condition, and one or more optional constraints (which must hold between the start and end of the goal). Our DSL ultimately grounds each goal to one or more objective and deterministic predicates over game states (e.g. interactions between entities, or between the player and the world), which allows in principle for the programs to be evaluated against traces of player behavior in *Minecraft*. This choice allows us to represent goals as verifiable programs, though we note that these programs are agnostic to the amount of player *effort* required to satisfy them. Our DSL does not structurally distinguish between a goal like "*craft a wooden sword*" and a goal like "*craft a diamond sword*," though the former is likely to be accomplished by any player and the latter must be decomposed into many intermediate steps and is a challenging task for even sophisticated artificial agents (Wang et al., 2023). In addition, our DSL does not readily facilitate translation of every goal – even conceptually simple goals like "*build a house*" cannot easily be rendered in this DSL because *Minecraft* does not provide an objective definition of *house*. Of the 150 goals selected for rating, 100 could be translated into our DSL.

[1] Participants in both generation and rating tasks were paid $6, and both tasks took slightly less than half an hour on average.

[2] Each scale contains the options "Not at all X," "Not very X,", "Somewhat X", "Moderately X," and "Very X"

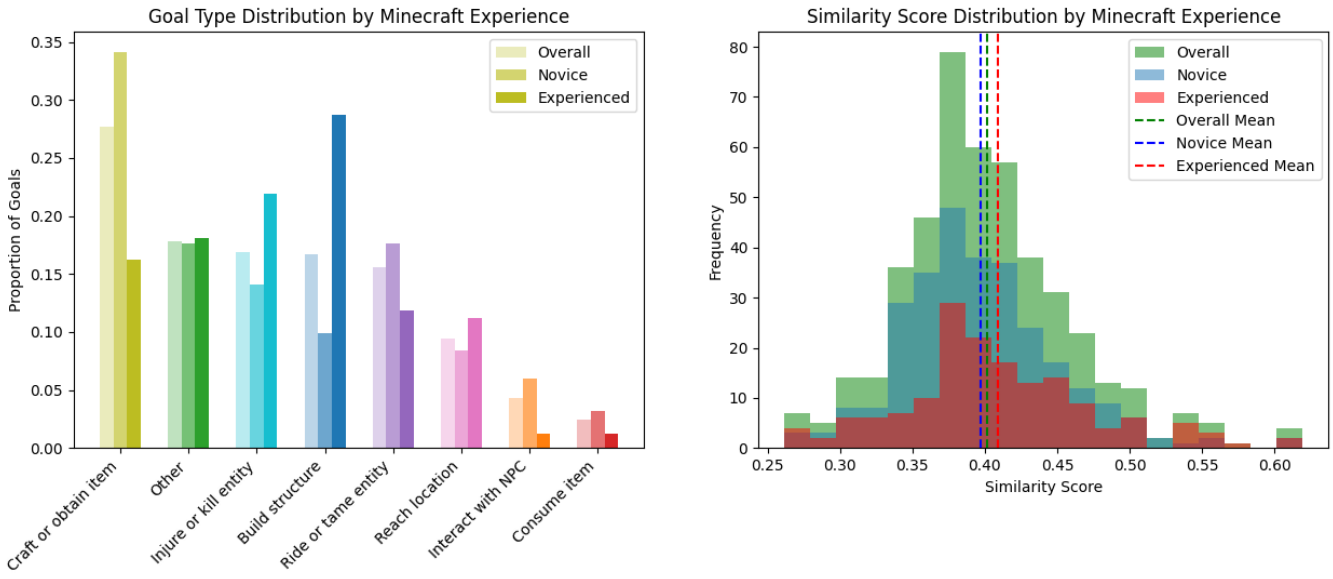

Figure 2: **(Left) Participants with different levels of Minecraft experience created different types of goals**. Human-created goals spanned a wide range of activities, from cliff-diving and explosive demolitions to combat with hostile enemies. We classified most goals into one of seven semantically distinct categories; about 18% were coded as "other". **(Right) Distribution of goal similarity scores to the *MineDojo* Reddit dataset, separated by creator experience level.** Goals span a wide range of similarity scores, with novice creators making goals that are on average less similar to the dataset than experienced creators.

Methods

There are many ways in which goals might differ, depending both on their content and their mode of representation (e.g. as natural language, formal program, or gesture). The *Minecraft* domain allows us to study goals that vary in a broad range of properties and implement analyses which might be difficult to apply in other less structured environments. Here we describe how each of our features are obtained, and in the following section we analyze their effects on ratings when grouped into semantically-related mixed-effects models.

Measuring Semantic Content

We begin by manually assigning each goal and its "complexified" version to one or more of 8 high-level semantic categories. These categories are: *craft or obtain an item*, *ride or tame an entity*, *injure or kill an entity*, *build a structure*, *interact with a non-player character*, *reach a location*, *consume an item*, and a final catch-all *other* category. Goals in the *other* category typically involve complex or uncommon activities, such as "*Use hoppers to automate the furnace*" or "*go X days without dying*." Different kinds of goals require different sorts of player actions, and we expect them to receive different ratings from evaluators.

Measuring Familiarity

We expect that the extent to which a novel goal appears expected or familiar plays a role in how it is evaluated and that one's domain expertise affects the kinds of goals they generate. Most obviously, we measure domain familiarity with the demographic *Minecraft* experience questions asked to participants in both the goal generation and goal rating tasks. We say that a creator or rater is "experienced" with *Minecraft* if they have played the game within the past year and "novice" otherwise (though we note that this categorization excludes participants who may have had extensive *Minecraft* experience in the past but not recently, and we highlight this as an important distinction for future study).

We also aim to capture the extent to which the *text* of each goal might be familiar to potential raters by leveraging an extant dataset of *Minecraft*-related discussion. Inspired by prior work which leverages the semantic embeddings of large neural networks to evaluate creativity (Beaty et al., 2022; Kandemirci et al., 2025), we use the `SentenceTransformers` Python library (Reimers & Gurevych, 2019) to analyze the *MineDojo* Reddit dataset, which contains more than 340,000 posts from the `r/Minecraft` subreddit (L. Fan et al., 2022). To match the length and general structure of the participant-created goals, we extract only the titles of the 50,684 text-only posts in the dataset and obtain their semantic embeddings using the `all-MiniLM-L6-v2` model. We then compute a "similarity score" for each goal by taking the average cosine similarity between its embedding and the embeddings of the k most similar post titles (setting $k = 506$ to equal 1% of the dataset size). What does this score measure? One account is that Reddit users are likely to talk about the things they find interesting, in which case the score would correlate positively with that measure. On the other hand, goals with high similarity scores may be more "expected" and thus less interesting. We explore these possible accounts below.

Measuring Concreteness

Some goals naturally afford multiple interpretations while others are more constrained, and we expect these differences to influence their evaluations. In *Minecraft*, the concreteness of a goal is in part determined by the game's hard-coded mechanics. For instance, "crafting" items in *Minecraft* involves arranging specific game objects in a particular pattern according to a fixed recipe. While a goal like "*craft a diamond sword*" might be temporally extended and decompose into multiple sub-goals, there is ultimately only one way to accomplish it. In contrast, a goal like "*build a giant treehouse high up in the trees*" offers variety both in potential *outcome* (i.e. what the resulting treehouse might look like) and in *policy* (i.e. what actions are necessary to construct the treehouse). To capture this kind of variation, we use the semantic categorization described above. Specifically, we consider the *crafting* and *building* categories to be opposite extremes of concreteness and treat whether a goal belongs to each as a pair of binary features. We also capture concreteness *linguistically* by using the word-level average of the norms collected by Brysbaert et al. (2014). These norms capture "the degree to which the concept [...] refers to a perceptible entity," which may in turn effect the variety of interpretations the goals afford. Finally, we can leverage our DSL to get another lens on concreteness by treating the ability to represent a goal programmatically itself as a feature. The general limitation of our DSL (and of any DSL which aims to be computationally evaluable) is that it cannot accommodate subjectivity. In this way, "translatability" serves to capture another view on concreteness.

Measuring Complexity

We anticipate that the effort needed to process or simulate a goal has some bearing on its evaluation, particularly with respect to its difficulty. As a basic validation of this hypothesis, we consider the "complexification" process described above. Participants in the goal generation task were explicitly instructed to make their goals more difficult and interesting, so we include a simple binary feature which tracks whether a goal has been "complexified."

We also consider the surface-level *linguistic complexity* of a goal based on its sentence length and syntactic complexity based on the maximal and average depth of its dependency parse tree (obtained with `spAcy` Python library). Longer sentences are in general more difficult to parse (Yasseri et al., 2012), while deeper syntax trees are more likely to require integrating information from across the sentence (Gibson, 1998). We also consider the *programmatic complexity* of a goal using its representation in our DSL. We collect a set of features analogous to the linguistic features described above: the size of the program parse tree, as well as its maximal and average depth. As our complexity measures depend on the existence of a programmatic representation, we restrict our analyses of this hypothesis to include only the 100 games that could be translated into our DSL.

What Kinds of Goals do People Make?

To begin, we explore the variety of goals generated by participants using our measures of semantic content and familiarity. In Figure 2 (left panel), we present the distribution of goals across the labeled goal types and broken down by the goal creator's *Minecraft* experience level. Notably, we see that *crafting* and *building* type goals exhibit the largest differences in prevalence between novices and experienced participants, with the former creating more crafting-based goals and the latter favoring the building-based goals. Indeed, these differences in goal type distribution between novice and experienced are significant overall ($\chi^2 = 47.379, p = 4.707 \times 10^{-8}$). Further, 83% of goals have at least one non-*other* goal type and no goal type captures more than 30% of goals, indicating that this semantic categorization meaningfully distinguishes between goals while still describing most of them.

In the right panel of Figure 2 we present the distribution of Reddit similarity scores for each goal, again broken down by creator domain experience. Participant-generated goals cover a wide range of scores (from 0.261 to 0.620), and qualitatively we see that high-similarity goals like "*trade with a villager*" almost exactly match multiple Reddit posts (e.g. "*Villager trading*" and "*Quick question about villager trading*") while low-similarity goals like "*ride an entity*" do not perfectly align with their nearest posts (e.g. "*Disappearing entities*" and "*Mysterious entity problem*"). Quantitatively, the distributions of similarity scores between novice and experienced goal creators are significantly distinct, with novice creators generating goals that are on average less similar to the extant corpus (Welch's t-test, $p = 0.049$).

Taken together, these results not only motivate the use of our particular features but also the use of *Minecraft* as a domain more generally. Owing to the fact that *Minecraft* exists as a game in the world and not just an environment in our experiment, participants naturally vary in their level of exposure and expertise. This variance, in turn, appears to affect both the kinds of goals that are generated and the extent to which they match other domain-related discourse. The similarity measure, too, is a benefit of using *Minecraft* – few games come with such an expansive corpus of text.

What Explains Goal Ratings?

In this section, we investigate how goal familiarity, concreteness, and complexity impact how people evaluate novel goals. To analyze these hypotheses, we use Cumulative Link Mixed Models (Christensen & Brockhoff, 2013) to predict the categorical rating variable as a function of one or more fixed effects, with random effects for the goal creator, specific goal, and goal rater. We z-score each continuously-valued feature (e.g., goal length) before performing the analysis, but keep binary-valued features (e.g., whether the goal has been "complexified") as-is.

Effects of Familiarity

Using the mixed-effects procedure described above, we consider four potential models of novel goal evaluation de-

Category	Model Name	AIC for Model of Rating				
		Interest	*Creativity*	*Fun*	*Difficulty*	*Teaching*
Familiarity	`null`	4340.748	4444.353	4205.585	**4514.585**	4267.737
	`experience`	4344.477	4447.119	4209.963	4517.436	4269.332
	`Reddit`	**4334.039**	**4439.866**	**4201.452**	4516.175	**4259.113**
	`full`	4337.769	4442.644	4205.921	4518.970	4260.892
Concreteness	`null`	4340.748	4444.353	4205.585	4514.585	4267.737
	`semantic`	**4313.193**	**4410.854**	4182.405	4512.717	**4239.514**
	`representation`	4333.403	4430.825	4204.653	**4506.996**	4264.738
	`full`	4315.095	4412.567	**4181.120**	4511.114	4240.010
Complexity	`null`	**2883.624**	2919.189	2764.016	3042.505	2721.426
	`linguistic`	2885.770	**2918.447**	**2762.757**	3040.969	**2714.557**
	`programmatic`	2887.568	2923.100	2768.086	**3037.622**	2723.205
	`full`	2889.570	2921.973	2766.327	3040.219	2717.711

Table 1: **Summary of Akaike information criterion (AIC) for mixed-effects models of goal ratings based on features of familiarity, concreteness, and complexity.** Model names describe semantically-linked groups of features. We highlight the best model within each hypothesis class and for each rating in bold.

rived from familiarity-based features. First, we consider a *null* model that includes only a single fixed effect for the "complexification" process (in addition to the random effects for goal creator, goal, and rater). We choose this as our null model to examine whether familiarity impacts ratings beyond the effects expected from explicitly modifying a goal to be more interesting and difficult. We then consider an *experience* model which additionally considers creator and rater domain experience as well as their interactions. Finally, we consider a *Reddit* model that includes a fixed effect for the Reddit similarity score of the goal, as well as a *full* model which combines the features of the base models.

We compare the models using the Akaike information criterion, presenting the AIC values in Table 1 (top third) and highlight the best model for each rating measure in bold. Somewhat surprisingly, our results indicate that the model which explicitly accounts for creator and rater domain experience does not best explain any of the rating measures, even failing to improve over the *null* model. Rather, it is the model that incorporates the similarity scores derived over the extant dataset which best explains each rating except for difficulty. In Table 2 (top third) we present the coefficients of the fixed effects for the best model on each rating measure (we omit the coefficients of *experience* model as it was never selected as the best). From this, we can see a clearer account of *what* the similarity scores capture. Coefficients across the board are positive and significant, indicating that the Reddit dataset likely captures specifically the topics that interlocutors find engaging and novel, rather than a high score indicating a goal that is expected and thus uninteresting.

Effects of Concreteness

To measure the effect of concreteness on goal evaluations, we once again consider four potential models. In addition the same *null* model as above, we implement a *semantic* model that adds fixed effects for the features based on semantic categorization and the average word-level concreteness norms. A simple *representation* model adds a fixed effect for the binary "DSL translability" feature, which is combined with the *semantic* model to form the *full* model. We present the AIC values of each model for each rating measure in Table 1 (middle third).

We find that relatively wide range of concreteness features are useful in explaining goal ratings. The *semantic* model best captures ratings of interest, creativity, and teaching; the *representation* model best captures ratings of difficulty; and the *full* model best captures ratings of fun. In each case, the *null* model is outperformed by one or more alternatives. Examining the coefficients in Table 2 (middle third), we see that whether a goal belongs to the *building* category is a strikingly significant factor with large positive coefficients on many ratings. This is somewhat unsurprising: building large free-form structures offers the ability for players to express themselves in an almost totally unconstrained manner. In addition, the coefficient for the "translability" feature on ratings of difficulty is negative, suggesting our DSL potentially excludes more challenging goal structures despite accounting for temporal extensiveness and compositionality. Moreso than any particular coefficient values, however, we see generally that the particulars of the game environment (i.e. differences between free-form building and recipe-based crafting) help explain goal evaluations. Interestingly, we see that the AIC values of the best concreteness-based models are also generally smaller than those of the best familiarity models. This may indicate that notions of specificity and concreteness are *particularly* salient to goal ratings, though a more dedicated likelihood ratio test would be required to make strong claims in this regard.

	Rating									
	Interest		*Creativity*		*Fun*		*Difficulty*		*Teaching*	
Feature	Coeff.	Sig	Coeff.	Sig	Coeff.	Sig	Coeff.	Sig	Coeff.	Sig
	(Reddit)		**(Reddit)**		**(Reddit)**		**(null)**		**(Reddit)**	
complexified	0.431 ±0.17	*	0.786 ±0.16	***	0.126 ±0.17	–	0.857 ±0.18	***	0.249 ±0.16	–
reddit_sim	0.288 ±0.10	**	0.230 ±0.09	*	0.238 ±0.10	*			0.306 ±0.09	**
	(semantic)		**(semantic)**		**(full)**		**(repr.)**		**(semantic)**	
complexified	0.341 ±0.16	*	0.725 ±0.15	***	0.065 ±0.16	–	0.779 ±0.18	***	0.193 ±0.16	–
is_building	1.360 ±0.26	***	1.412 ±0.23	***	1.562 ±0.30	***			1.494 ±0.25	***
is_crafting	−0.154 ±0.23	–	−0.077 ±0.20	–	−0.078 ±0.22	–			0.295 ±0.22	–
concreteness	−0.130 ±0.09	–	−0.079 ±0.08	–	−0.191 ±0.09	*			−0.056 ±0.09	–
has_program					0.454 ±0.25	–	−0.740 ±0.24	**		
	(null)		**(linguistic)**		**(linguistic)**		**(program)**		**(linguistic)**	
complexified	0.422 ±0.20	*	0.479 ±0.25	–	−0.185 ±0.26	–	0.595 ±0.24	*	−0.327 ±0.25	–
goal_length			0.427 ±0.27	–	0.367 ±0.28	–			0.847 ±0.28	**
syntax_max			0.080 ±0.42	–	0.919 ±0.46	*			0.081 ±0.42	–
syntax_avg			−0.078 ±0.47	–	−0.906 ±0.51	–			−0.309 ±0.47	–
program_tree							−0.044 ±0.20	–		
program_max							0.179 ±0.65	–		
program_avg							0.877 ±0.53	–		

Table 2: **Coefficient summary for mixed-effects models of goal ratings based on features of familiarity, concreteness, and complexity (vertical groups).** The headers of each column indicate which model within each category (see Table 1) performed best on that rating measure based on AIC, which are then broken down by the individual features they include.

Effects of Complexity

Finally, we turn to the effects of complexity. Here, our *null* model with a single fixed effect for the "complexification" process most obviously captures the "background features" against which we wish to compare. We consider a *linguistic* model which adds fixed effects for goal length and syntax tree depth (maximum and average) and a *programmatic* model which analogously contains fixed effects for program tree size and depth (maximum and average). The *full* model, as before, combines the two sub-models. We note again that in order to ensure a fair comparison between linguistic and programmatic features, we consider only the 100 rated games for which we were able to obtain a translation into our DSL. As such, the AIC values for the complexity-based models should not be compared directly to those in other categories.

With that in mind, we examine the values in Table 1 (bottom-third). We see that the *linguistic* model best explains ratings of creativity, fun, and amount taught, while the *programmatic* model best explains ratings of difficulty and the *null* model outperforms the alternatives on ratings of interest. From this we can conclude that measures of complexity, variously defined, *do* help explain ratings of goals. Further, a programmatic representation which explicitly abstracts over linguistic content is particularly helpful in explaining ratings of goal difficulty. We note the resonance with our similar finding on concreteness: the "translatability" of a goal into a programmatic representation helps explains ratings of difficulty, and the features of those programs help explain ratings of difficulty for the games over which they are defined. The specific programmatic features selected are likely less important than the fact that the programmatic representation in general is a useful predictive tool.

Discussion

Taken together, we find evidence to support each of our hypotheses: ratings of novel goals are explained in part by quantifiable measures of familiarity, concreteness, and complexity. Different features are better at explaining different rating measures: syntactic and semantic features are useful for modeling ratings of creativity, fun, and how much a goal teaches, while features based on the programmatic representation of a goal in a domain-specific language are particularly useful for explaining ratings of difficulty. Across a variety of measures, features derived from semantic categorization and an extant domain-related corpus are particularly predictive. While our results are specific to a single domain, they show how structured game environments in general and *Minecraft* in specific are useful for understanding the broad process of human goal creation. Future work could fruitfully evaluate goal feasibility through automated play and study the effects of multi-player interactions.

Games can be powerful tools for understanding a wide range of human behavior and cognitive capacities (Allen et al., 2024), and here we demonstrate their utility in studying the fundamental process by which humans invent new tasks and objectives for themselves.

Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant DGE-2234660. The authors would like to thank the NLP team at Microsoft Research Redmond for their advice and discussions.

References

Allen, K., Brändle, F., Botvinick, M., Fan, J. E., Gershman, S. J., Gopnik, A., ... others (2024). Using games to understand the mind. *Nature Human Behaviour*, 1–9.

Beaty, R. E., Johnson, D. R., Zeitlen, D. C., & Forthmann, B. (2022). Semantic distance and the alternate uses task: Recommendations for reliable automated assessment of originality. *Creativity Research Journal*, *34*(3), 245–260.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, *46*, 904–911.

Canossa, A., Martinez, J. B., & Togelius, J. (2013). Give me a reason to dig minecraft and psychology of motivation. In *2013 ieee conference on computational inteligence in games (cig)* (pp. 1–8).

Christensen, R. H. B., & Brockhoff, P. B. (2013). Analysis of sensory ratings data with cumulative link models. *Journal de la Société Française de Statistique*, *154*(3), 58–79.

Chu, J., Hu, J., & Ullman, T. D. (2024). The task task: Creative problem generation in humans and language models. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Chu, J., Tenenbaum, J. B., & Schulz, L. E. (2023). In praise of folly: flexible goals and human cognition. *Trends in Cognitive Sciences*.

Davidson, G., Gureckis, T. M., & Lake, B. (2022). Creativity, compositionality, and common sense in human goal generation. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 44).

Davidson, G., Todd, G., Togelius, J., Gureckis, T. M., & Lake, B. M. (2025, May). Goals as Reward-Producing Programs. *Nature Machine Intelligence*. Retrieved from https://www.nature.com/articles/s42256-025-00981-4

Dweck, C. S. (1992). Article commentary: The study of goals in psychology. *Psychological Science*, *3*(3), 165–167.

Elliot, A. J., & Fryer, J. W. (2008). The goal construct in psychology. *Handbook of motivation science*, *18*, 235–250.

Fan, L., Wang, G., Jiang, Y., Mandlekar, A., Yang, Y., Zhu, H., ... Anandkumar, A. (2022). Minedojo: Building open-ended embodied agents with internet-scale knowledge. In *Thirty-sixth conference on neural information processing systems datasets and benchmarks track*. Retrieved from https://openreview.net/forum?id=rc8o_j8I8PX

Fan, Y., Lane, H. C., & Delialioğlu, Ö. (2022). Open-ended tasks promote creativity in minecraft. *Educational Technology & Society*, *25*(2), 105–116.

Fishbach, A., & Ferguson, M. J. (2007). The goal construct in social psychology. *Social psychology: Handbook of basic principles*, *2*, 490–515.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76.

Kandemirci, B., Beaty, R. E., Johnson, D., Oliver, B. R., Kovas, Y., & Toivainen, T. (2025). What is creative in childhood writing? computationally measured linguistic characteristics explain much of the variance in subjective human-rated creativity scores. *Learning and Individual Differences*, *118*, 102626.

Lane, H. C., & Yi, S. (2017). Playing with virtual blocks: Minecraft as a learning environment for practice and research. In *Cognitive development in digital contexts* (pp. 145–166). Elsevier.

Mojang Studios. (2011). *Minecraft*. Mojang Studios. Retrieved from https://www.minecraft.net

Molinaro, G., & Collins, A. G. (2023). A goal-centric outlook on learning. *Trends in Cognitive Sciences*.

Reimers, N., & Gurevych, I. (2019, 11). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing*. Association for Computational Linguistics. Retrieved from https://arxiv.org/abs/1908.10084

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE transactions on autonomous mental development*, *2*(3), 230–247.

Shaw, A. (2023). Creative minecrafters: Cognitive and personality determinants of creativity, novelty, and usefulness in minecraft. *Psychology of Aesthetics, Creativity, and the Arts*, *17*(1), 106.

Sim, Z. L., & Xu, F. (2019). Another look at looking time: Surprise as rational statistical inference. *Topics in cognitive science*, *11*(1), 154–163.

Voiskounsky, A. E., Yermolova, T. D., Yagolkovskiy, S. R., & Khromova, V. M. (2017). Creativity in online gaming: Individual and dyadic performance in minecraft. *Psychology in Russia*, *10*(4), 40.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., ... Anandkumar, A. (2023). Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Willatts, P. (1999). Development of means–end behavior in young infants: pulling a support to retrieve a distant object. *Developmental psychology*, *35*(3), 651.

Yasseri, T., Kornai, A., & Kertész, J. (2012). A practical approach to language complexity: a wikipedia case study. *PloS one*, *7*(11), e48386.