User Study: Automapper comparison

Chrisvenator

2025-03-20

Since our English is not good, we used ChatGPT and Grammarly to help formulate sentences better!

User study analysis

In this analysis, we will compare how well different automappers perform. For this, we asked 32 participants to rate maps from 1 (worst map possible) to 10 (best map you have ever played). There were 4 maps made by BeatKenja, 1 by BeatSage, 3 by BeatStormer, 2 by InfernoSaber, 3 Human maps and 2 Hybrid maps.

- The human maps were a control group to see if the participants had any bias towards AI-generated maps.
- The skeleton of the two hybrid maps was created by BeatKenja but then heavily cleaned and remapped by a human. It is already to be considered human because there is nothing left that was generated.
- The maps made by the Automappers BeatKenja, BeatSage, BeatStormer, and InfernoSaber were left as is and not cherry picked. They were not modified in any way to simulate how a regular user would use them.

Used maps:

A player's map preferences is like a music taste. It's different for everyone. And everyone loves and hates certain maps. So we prepared multiple human maps from different playstyles to account for that.

In this table, we define a "good player" as a player who is comfortable with 12 stars or more. You would find a player like this typically in the top 500 worldwide. But not everyone plays ranked maps so there are players that are in the top 3000 who would be comfortable with that. More details are explained later.

We asked participants to rate the following maps:

Map	Mapper	Comment
Hedonism	Human	The skeleton of this map was created by BeatKenja. But it has been edited so much that it can not be considered an AI map anymore. This is a mid-speed and a slight tech map. It's fast but not too fast and has some tech angles in it. Good players should be able to easily get a score of 96% on this.
Requiem	Human	The skeleton of this map was created by BeatKenja. But it has been edited so much that it can not be considered an AI map anymore. It can be considered as a slightly slower midspeed map. In addition to that, it is short to see if the length has any effect on the rating. Good players should be able to get a score of 97% on this.
Haru-machi Clover		
Gravity Falls	BeatKenja	This map is not very fast but has some weird angles here and there which could make this map interesting. A good player should be able to get 97% on it.

Map	Mapper	Comment
Rolling Girl	Human	This map is considered a tech map. There are a lot of players who love and who hate tech maps. This map ensures that we cover these players. A good tech player is expected to get a 95% on it. Meanwhile, a good speed player might only get 80% to 90%. We chose this map in order to see if players would give this map a bad review just because it is a tech map.
Rolling Girl	BeatKenja	This map was created by taking the timings from map above (Rolling Girl mapped by Alice) and generating a map based on the timings. No modifications apart from lighting have bee done. This map was chosen so that we could see if good timings make a difference. Another quirk of this map is that it is quite fast and has a lot of inlines, staircases and windows. Only problem is that thesequirks might not fit to the beat.
Jeff Bezos	BeatKenja	This map is considered an acc map. It is very slow and players focus on hitting the notes perfectly to get as much points as possible. It is common that good players get about 99% or more on maps like this.
P*Light - SAtAN	Human	This map should represent a speed map. It is very fast without breaks but the notes are very predictable. We chose the Hard difficulty because we expected that not a lot of participants would be able to play the upper difficulties. We chose this map becuase it is popular and well rated on beatsaver with 1068 upvotes and 207 downvotes as of March 2025.
Lusumi - Out of this Planet	Human	This map is almost a hell-tech map. We have asked for recommendations of good tech maps and this one was recommended quite a lot. It has good timings, lots of angles and fast but fun transitions. It has excellent flow and the patterns are predictable but not repetetive. It has it's own quirk which is consistent throughout the entire map. It is also has a few fitbeat elemnts in the middle. We chose this map as a control map to see if players would downvote a map just because it is tech. The map description even says: "Dont blame me if you miss cause its a skill issue". A good tech player is expected to get between 94% and 96%. A non-tech player will struggle to beat it.
LittleVMills - Gurenge	BeatKenja	We used the timings from the map "5270" mapped by DigiRago to see if good timings make a difference. This map is slower than the rest before. It's more of an acc map. A good player should get about 96.5%. Pro players will be able to get about 98%. Since there are two "LittleVMills - Gurenge", we will call this map "LittleVMills... Gurenge.BK"
LittleVMills - Gurenge	BeatSage	We included this map to show how BeatSage stands up against another automapper. The map has a lot of resets, DDs, weird angles and in general very bad timings. A good player could get 98% on it. But because of all these mapping errors, the scores will vary. If players really feel like getting a good score, they will be able to do it regardless of rank. But probably not many will want to. We expect a lot of participants to abort this level. Since there are two "LittleVMills - Gurenge", we will call this map "LittleVMills... Gurenge.BS"
Malefisole	BeatStorm	The expert+ version is very overmapped and probably not enjoyable because of it's unpredictable patterns. So we requested everyone to play the difficulty expert. But not many people did read the task so we will be analyzing the ex+ difficulty. From good players we expect a score of about 92% to 94%. If this were a ranked map, we could expect a 97%.

Map	Mapper	Comment
Requiem Haru-machi	BeatStormer	With this map, we canted to see the difference between BeatStormer and a Human mapper. How the scores and ratings differ. It starts out as a balanced midspeed map. But in the middle, the patterns are changing to sideway inlines which are difficult to hit, when not set up properly. A good player should get 95% if this were ranked. Since it is not, we expect a good player to be between 90% and 93% with a FC.
Tetoris	BeatStormer	This map has decent/good timings for an AI and the patterns are varied and could be very interesting. On the other hand, it has a lot of unpredictable resets after sideway notes. This map could be interesting to tech players without the random resets. We expect good players to get a 96% on it. But because of the many resets, probably a lot of players will abort it. Since there are two maps called Tetoris, we will call this one "Tetoris.BST" (BeatStormer)
Tetoris	Infernosaber	The timings are very off-beat. Patterns are simple but too fast for this map. It has a lot of uncomfortable inward jumps. A good player will be able to get 95% FC on this. But we expect that many players will abort it. Since there are two maps called Tetoris, we will call this one "Tetoris.IS" (Internosaber)
They will not escape	Infernosaber	The patterns of this map are very predictable and there are not many parity breaks. The timings are also decent for an AI. On the other hand, there are a lot of unexpected transitions and patterns which makes this map uncomfortable to play for a lot of players. A good player should get 94%.

Here is additional information about each map.

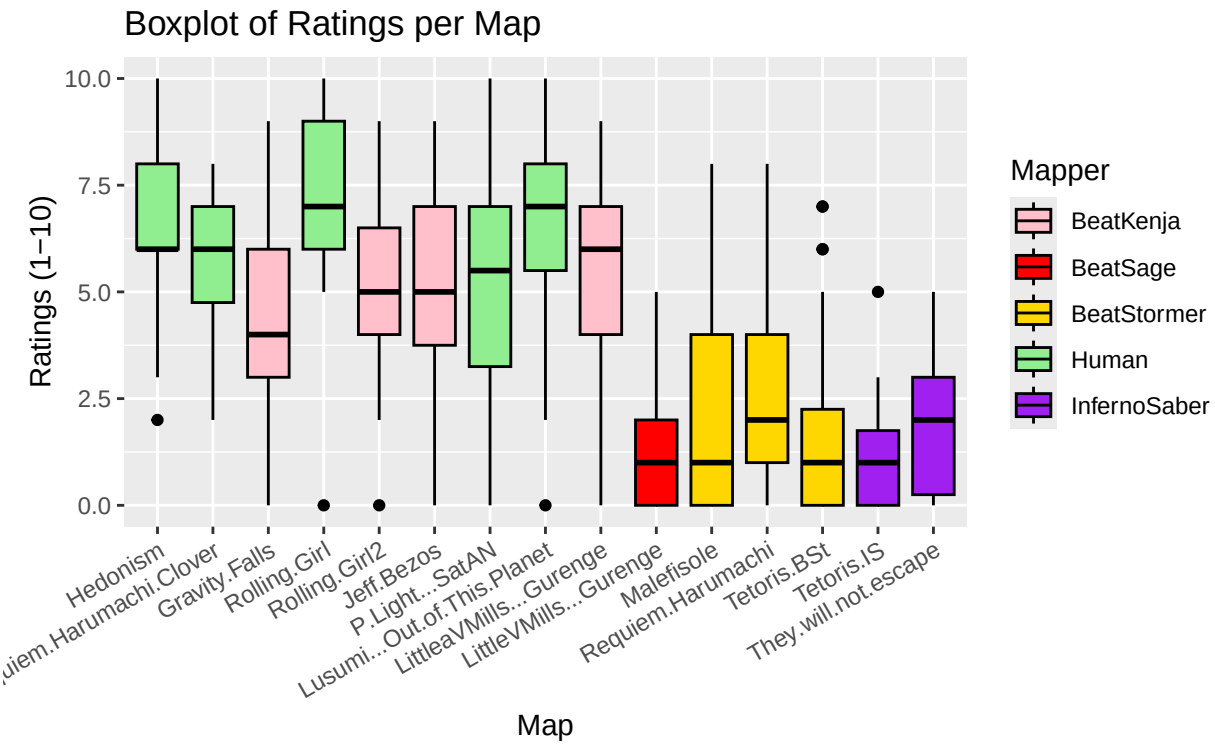
Map	Difficulty	Mapper	nps	Estimated Stars	ID
Hedonism	Ex+	Human	7.13	5.76	32d8d
Requiem Harumachi Clover	Ex	Human	5.87	4.68	31adf
Gravity Falls	H	BeatKenja	5.98	3.55	43fcb
Rolling Girl	Ex+	Human	6.21	7.12	3024a
Rolling Girl	Ex+	BeatKenja	6.21	4.58	44212
Jeff Bezos	N	BeatKenja	3.55	0.91	3dec2
P*Light - SAtAN	H	Human	8.77	8.77	13735
Lusumi - Out of this Planet	Ex+	Human	9.01	7.41	39d0e
LittleVMills - Gurenge	Ex+	BeatKenja	5.57	3.68	43fcc
LittleVMills - Gurenge	Ex+	BeatSage	3.81	3.34	43fcd
Malefisole	Ex+	BeatStormer	7.27	8.91	4420c
Requiem Harumachi	Ex+	BeatStormer	8.29	8.11	4420e
Tetoris	Ex+	BeatStormer	6.89	7.79	4420f
Tetoris	Ex+	Infernosaber	10.11	10.96	44210
They will not escape	Ex+	Infernosaber	8.58	7.15	44211

The stars were estimated by the Chromapper Plungin "Error Checker": Link

E - Easy
N - Normal
H - Hard
Ex - Expert
Ex+ - Expert+

Boxplot of the ratings

This boxplot displays the ratings given to different Beat Saber maps, categorized by their mapper type:



Human-made maps generally received higher ratings Maps like `Hedonism`, `Rolling Girl (Human)`, and `Lusumi - Out of this Planet` received high median ratings (~7-8). These maps also have a wider spread, indicating that different players had different opinions about them.

Auto-generated maps tend to have lower ratings Maps from InfernoSaber (`Tetoris`, `They will not escape`) and BeatStormer (`Requiem Harumachi`, `Malefisole`, `Tetoris BSt`) had lower median ratings (~2-5). This suggests that automapper-generated maps may be perceived as lower quality or less enjoyable compared to human-made maps.

BeatKenja's maps have mixed reactions `Rolling Girl (BeatKenja)` and `Gravity Falls` have a wide range of ratings. Some players liked them (~6-8 range), while others rated them poorly (~2-4). This suggests that BeatKenja's automaps may be enjoyable for some but not for everyone.

BeatSage received one of the lowest ratings `LittleVMills - Gurenge (BeatSage)` has a very low median (~2-3) and a narrow spread. This means most players gave it a low score consistently. Since the same song mapped by BeatKenja received a higher rating, this suggests that BeatSage's maps might be less enjoyable.

Tech maps have polarizing opinions `Rolling Girl (Human)` and `Lusumi - Out of this Planet` are tech maps (maps with complex angles and unconventional patterns). These maps have a wide spread, meaning some players love them while others dislike them. This aligns with the idea that tech maps have a "love it or hate it" effect.

Acc maps have Accurate timings & predictable patterns and might be rated higher like `Jeff Bezos (BeatKenja)`. These maps received higher ratings, likely because for players with acc maps, there does not need to be good flow and players have enough time to prepare for parity breaks since the map is so slow.

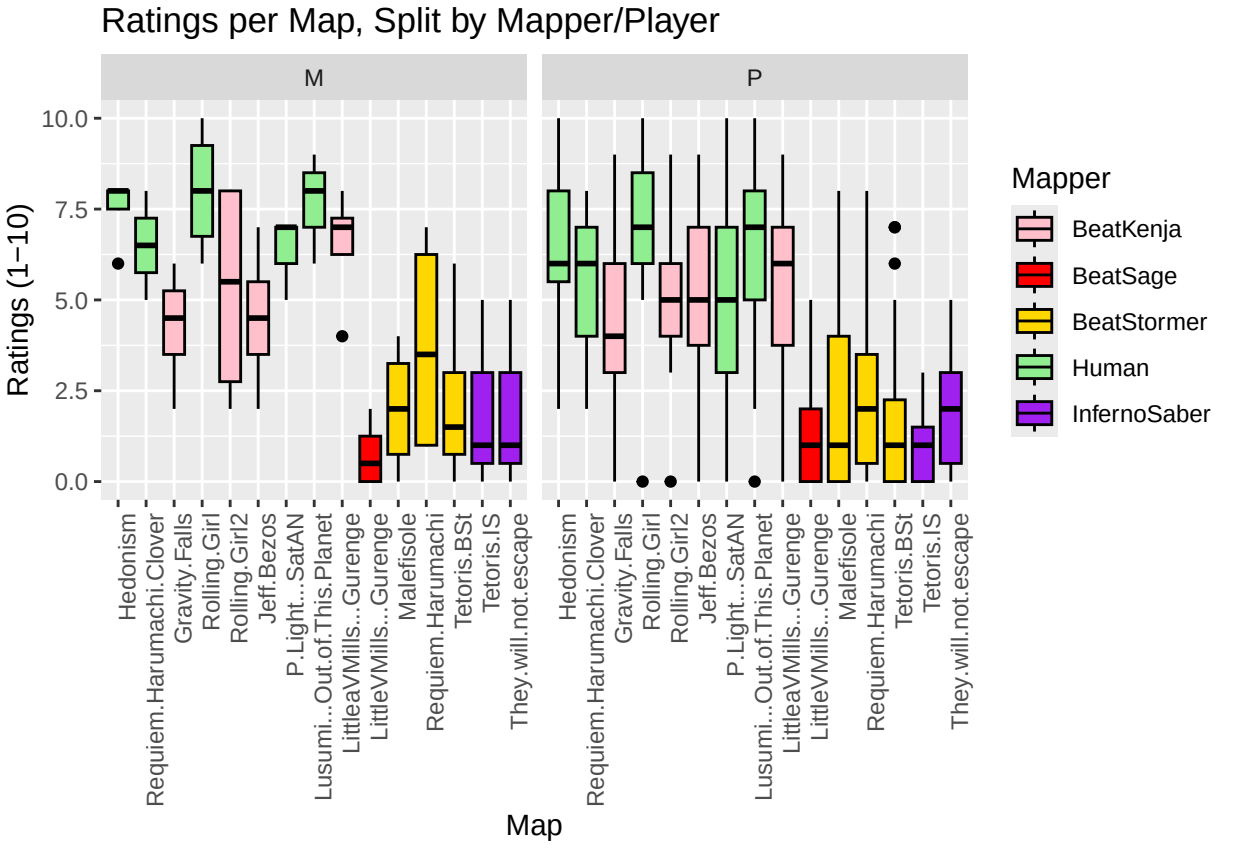
Player vs Mapper

In this section, we investigate whether Mappers and Players evaluate maps differently. The underlying assumption is that Mappers, due to their experience in level design, may be more critical and better at identifying structural flaws, whereas Players might judge maps based on subjective enjoyment and playability. This comparison aims to show whether these two groups have distinct evaluation patterns.

Participants in the user study were categorized into two groups:

1. **Mappers (M)**: Participants with direct experience in map creation.
2. **Players (P)**: Individuals who with no or little mapping experience.

To analyze potential differences in rating behavior, the ratings were split based on these two groups. The boxplots in the following Figure visualize how each group rated different maps. Notably, **only four mappers participated in the study**, which should be considered when interpreting the results!



Mappers tend to assign lower ratings than Players
Across most maps, the median ratings from Mappers are lower compared to those from Players. This supports the assumption that Mappers apply more critical evaluation criteria. Mappers likely focus on mapping techniques, note flow, and structural consistency, while Players prioritize playability and subjective enjoyment.

Higher agreement on Human-made maps
Maps created by human mappers (e.g., `Hedonism`, `Requiem Harumachi Clover`, `Lusumi - Out of This Planet`) received consistently higher ratings from both groups. This suggests that both Mappers and Players recognize the quality of well-crafted human maps. This is supported by the reduced variance in ratings for these maps which indicates a consensus regarding their playability and design consistency.

Lower ratings for AI-generated maps, especially among Mappers
Maps generated by InfernoSaber, BeatStormer, and BeatSage (e.g., `Tetoris, They Will Not Escape, Malefisole`) received significantly lower ratings from Mappers. While Players also rated these maps lower, the drop in ratings was more pronounced among Mappers. This suggests that Mappers may detect issues in AI-generated maps more easily, such as awkward note placements, lack of pattern structure, or timing inconsistencies.

Tech maps are divisive among Players, but Mappers rate them more consistently
`Rolling Girl (Human)` and `Lusumi - Out of This Planet` are technical maps with complex angles and unconventional note patterns. Among Players, ratings for these maps have a wider distribution, suggesting a love-it-or-hate-it dynamic. Mappers, however, rate these maps more consistently, likely evaluating them based on technical execution rather than personal preference.

Study Limitations

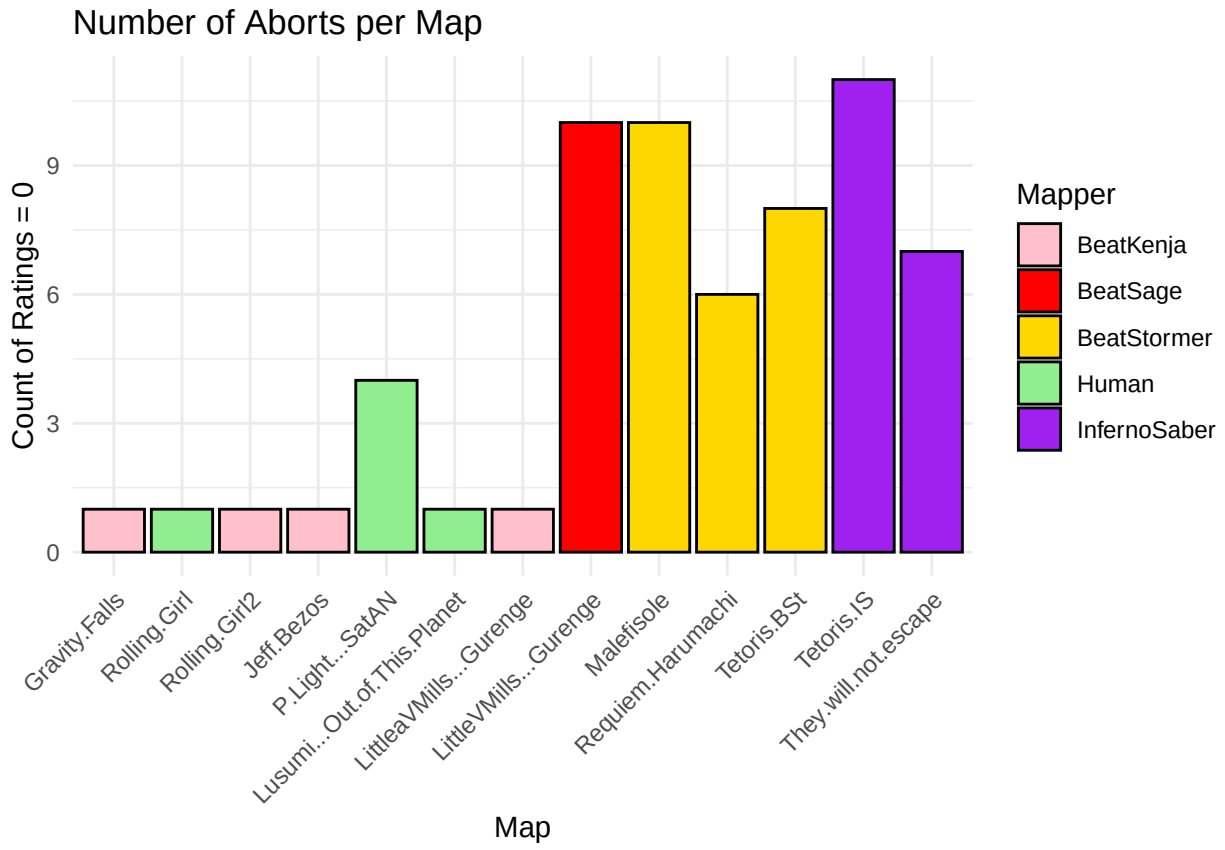
The small number of Mappers (n = 4) introduces potential bias, as their individual preferences may have had a significant influence on the results. Future studies should aim to increase the number of participating Mappers to obtain a more statistically robust comparison. Additionally, conducting statistical tests (e.g., t-tests, ANOVA) could determine whether the observed differences are statistically significant.

Conclusion

This comparison suggests that **Mappers apply stricter evaluation criteria than Players** and are more likely to **identify structural mapping flaws**. Players, on the other hand, display **greater variance in their ratings**, particularly for **tech maps**, where personal preference plays a significant role. The **lower ratings for AI-generated maps among Mappers** indicate that current automappers may not yet achieve the same level of quality as human-designed maps.

Analysis of Aborted Maps

This analysis investigates the number of abortions (early exits) per map, as visualized in the Figure below. An aborted map indicates that a player quit before completing it, which may suggest issues with difficulty, discomfort, or general lack of engagement.



Higher abort rates in AI-generated maps
The highest abort rates appear in InfernoSaber (`Tetoris.IS, They Will Not Escape`) and BeatStormer (`Malefisole, Requiem Harumachi, Tetoris.BST`). This suggests that automapper-generated maps may be frustrating or unplayable for some players, possibly due to bad flow, awkward note placement, or off-beat patterns. Tetoris.IS has the highest number of abortions, reinforcing the assumption that InfernoSaber's mapping may not be well-received.

Human-made maps have significantly lower abort rates
Maps by human mappers (`Gravity Falls, Rolling Girl, Hedonism, Jeff Bezos`) have very few aborts, indicating that players were more likely to finish them. The only human map with noticeable aborts is `P*Light - SatAN`, which is a fast speed map, likely leading to player fatigue or inability to keep up.

BeatSage-generated map (`LittleVMills - Gurenge`) also has a high abort rate
This aligns with previous findings that BeatSage maps received lower ratings, suggesting that players not only disliked playing them but also quit mid-session.

Tech-heavy maps
`Rolling Girl (Human)`, which is a tech map, has low abort rates, indicating that players were willing to engage with it despite its complexity. Even though `Lusumi - Out of This Planet`, another tech-heavy map, is a lot more difficult, it does not have a higher number of abortions, suggesting that difficulty does not matter as much in Tech maps as it does in speed maps.

Possible Explanations for High Abort Rates
Poor flow and structure in AI-generated maps: InfernoSaber and BeatStormer maps show high dropout rates, likely due to unpredictable patterns or awkward transitions. Extreme difficulty: Some maps may have been too difficult, particularly `P*Light - SatAN` and `Tetoris IS.`Frustration from bad mapping quality: BeatSage's `LittleVMills - Gurenge` had a high number of quits, supporting the claim that BeatSage-generated maps are not enjoyable.
Players also frequently mentioned issues such as:

- Frequent resets and uncomfortable angles, leading to frustration and quits.
- Poor note timing and lack of musical representation, making some AI-generated maps feel disconnected from the song's rhythm.
- Vision blocks and stacked notes, causing confusion and interrupting gameplay flow.

These comments directly correlate with the abort rates, particularly for InfernoSaber and BeatStormer maps, which received many complaints about randomness and poor pattern consistency.

Conclusion

The data suggests a strong relationship between abort rates and gameplay issues. Human maps generally provide better experiences, while InfernoSaber, BeatStormer, and BeatSage maps exhibit significant design flaws. Future research could explore whether AI improvements (e.g., reinforcement learning) can enhance map quality and reduce the problems mentioned above to reduce abort rates.

Playtime & rank

In this chapter we wanted to know if playtime or comfortable stars influence your rank and thereby influence your ratings. We have put the interpretation before the graphs to let each graph + summary be on the same page.

```
model1 <- lm(comfortable.with.XX.stars ~ Playtime, data = data)
model2 <- lm(Current.Rank.SS.BL ~ Playtime, data = data)
model3 <- lm(Current.Rank.SS.BL ~ comfortable.with.XX.stars, data = data)
```

```
##                              Model   R_Squared      P_Value
## 1     Playtime → Comfortable Stars 0.007850329 6.539032e-01
## 2          Playtime → Current Rank 0.008750683 6.358802e-01
## 3 Comfortable Stars → Current Rank 0.493832508 3.051979e-05
```

1. Influence of Playtime on Comfortable Star Level

This analysis examines whether total playtime in Beat Saber predicts the highest star difficulty a player feels comfortable with. A linear regression model (`lm(comfortable.with.XX.stars ~ Playtime)`) was applied to assess the relationship.

The results indicate **no significant correlation** between Playtime and Comfortable Stars ($p = 0.654$, $R^2 = 0.00785$). The regression coefficient (1.082e-04) suggests that each additional hour of playtime increases the comfortable star level by only 0.0001 stars, which is negligible. The flat regression line and wide confidence interval further support the weak relationship. This suggests that **skill progression is not directly tied to playtime** but may depend on other things. Future studies could explore explanations as to why that is.

2. Influence of Playtime on Rank

This analysis examines whether total playtime correlates with a player's rank. A linear regression model (`lm(Current.Rank.SS.BL ~ Playtime)`) was applied to assess this relationship.

The results show **no statistically significant correlation between Playtime and Rank** ($p = 0.636$, $R^2 = 0.00875$). The estimated effect size (9.824e-02) is minimal, suggesting that additional playtime does not strongly predict rank improvements. The flat regression trend and wide confidence interval further indicate high variance in rank progression. These findings suggest that **factors beyond playtime may play a larger role** in determining a player's rank.

3 Influence of Comfortable Star Level on Rank

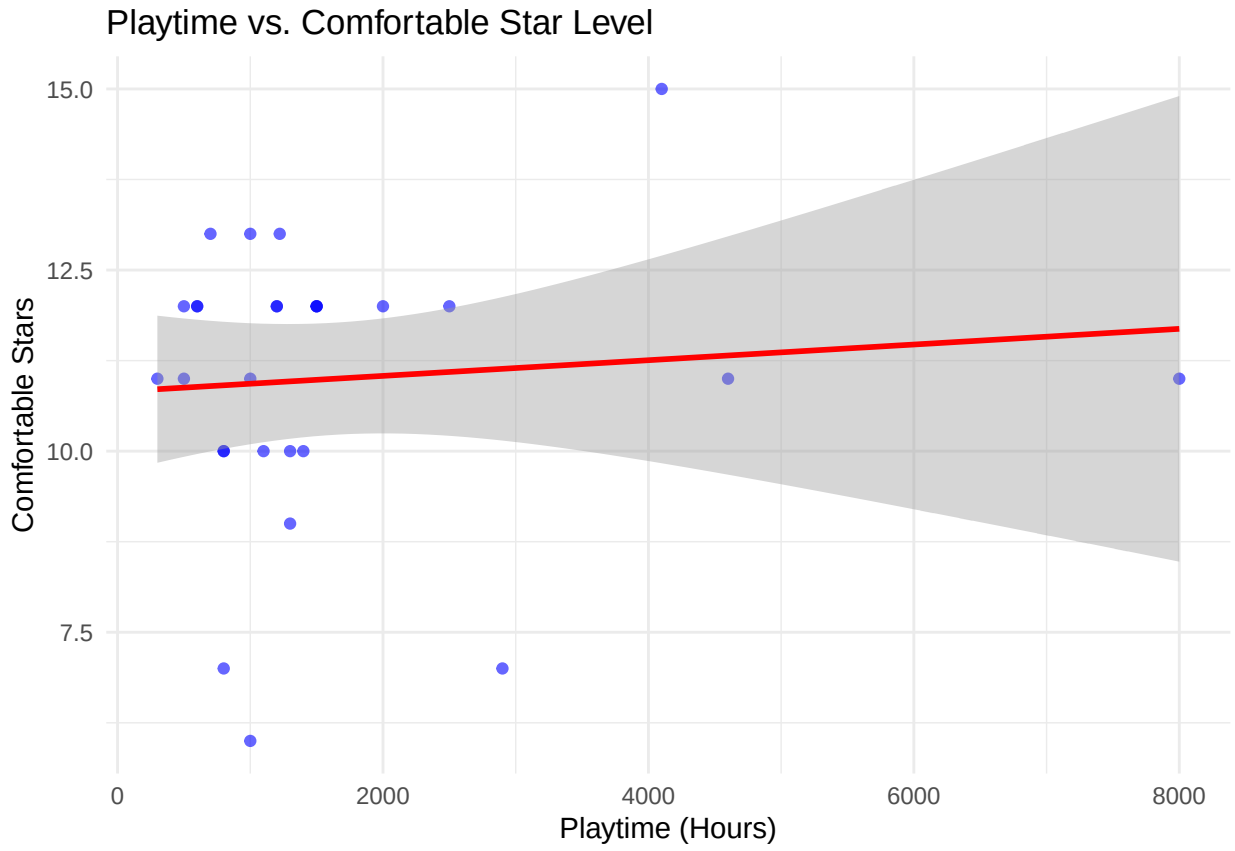
This analysis examines whether the highest star difficulty a player feels comfortable with correlates with their rank. A linear regression model (`lm(Current.Rank.SS.BL ~ comfortable.with.XX.stars)`) was applied.

The results indicate a **strong negative correlation** between Comfortable Stars and Rank ($\beta = -604.2$, $p < 0.001$, $R^2 = 0.4938$), meaning that for each additional star in a player's comfort level, their rank improves by approximately 604 positions. The statistically significant p-value (3.05e-05) and high R^2 value (49.38%) suggest that Comfortable Star Level is a **strong predictor of player rank**.

These findings indicate that comfortable star level is a better indicator of competitive rank than total playtime, reinforcing the idea that skill progression is tied more to difficulty adaptation rather than sheer hours played.

1. Does your playtime influence with how many stars you are comfortable with?

```
## `geom_smooth()` using formula = 'y ~ x'
```

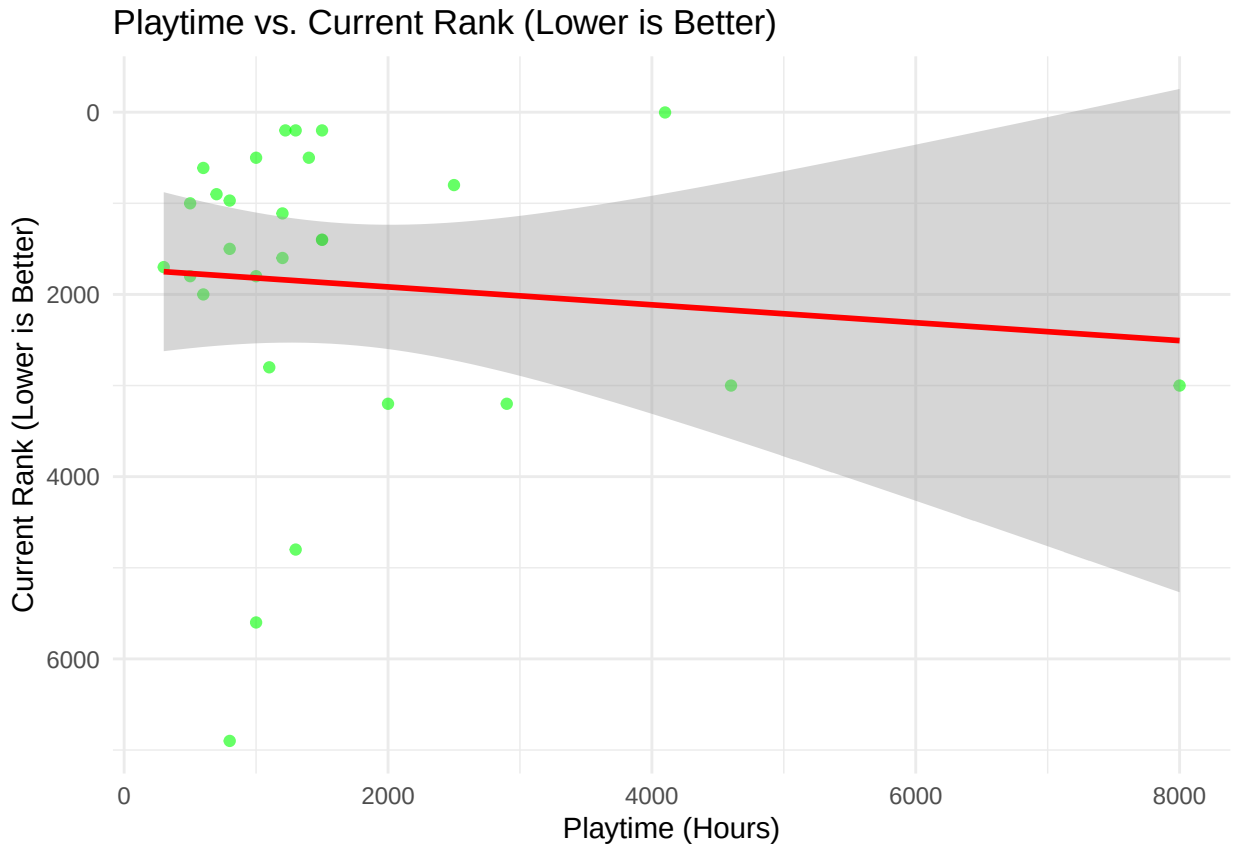



```
summary(model1)
```

```
##
## Call:
## lm(formula = comfortable.with.XX.stars ~ Playtime, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9307 -0.9172  0.5260  1.0639  3.7338
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.082e+01  5.430e-01  19.930   <2e-16 ***
## Playtime    1.082e-04  2.386e-04   0.454    0.654
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.992 on 26 degrees of freedom
## Multiple R-squared:  0.00785,    Adjusted R-squared:  -0.03031
## F-statistic: 0.2057 on 1 and 26 DF,  p-value: 0.6539
```

2. Does your playtime influence your rank?

```
## `geom_smooth()` using formula = 'y ~ x'
```

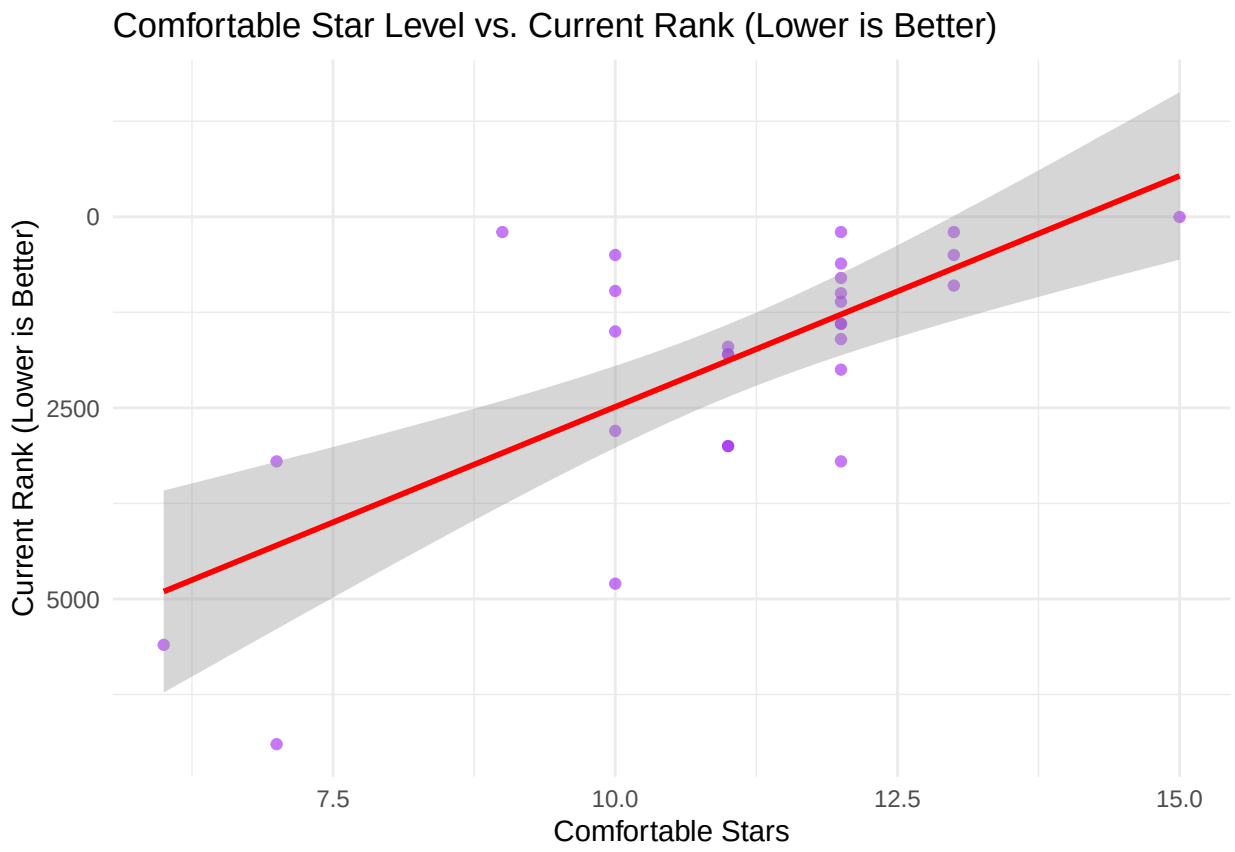



```
summary(model2)
```

```
##
## Call:
## lm(formula = Current.Rank.SS.BL ~ Playtime, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2120.7 -1166.8  -383.9   576.7  5100.5
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.721e+03  4.667e+02   3.688  0.00105 **
## Playtime    9.824e-02  2.051e-01   0.479  0.63588
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1712 on 26 degrees of freedom
## Multiple R-squared:  0.008751,   Adjusted R-squared:  -0.02937
## F-statistic: 0.2295 on 1 and 26 DF,  p-value: 0.6359
```

Does your comfortable stars influence your rank?

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
summary(model3)
```

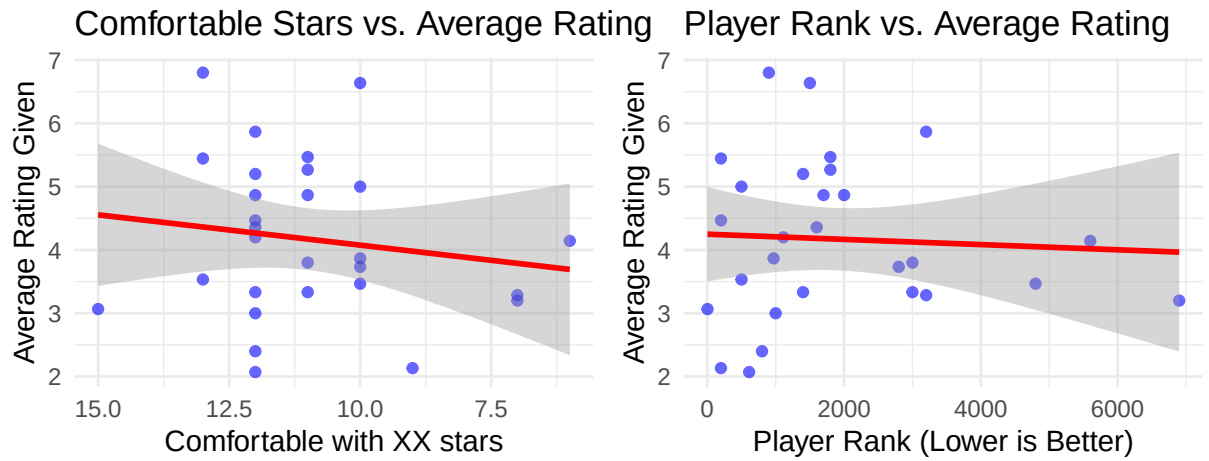
```
##
## Call:
## lm(formula = Current.Rank.SS.BL ~ comfortable.with.XX.stars,
##     data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2890.4  -524.8   -82.0   577.6  2601.3
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 8528.0     1339.7   6.366 9.65e-07 ***
## comfortable.with.XX.stars   -604.2      120.0  -5.037 3.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1223 on 26 degrees of freedom
## Multiple R-squared:  0.4938, Adjusted R-squared:  0.4744
## F-statistic: 25.37 on 1 and 26 DF,  p-value: 3.052e-05
```

Skill vs Rating given

Since comfortable star level is correlated with rank, this section examines whether player skill level (comfortable stars, rank) affects how they rate maps. Two linear regression models were tested:

```
model1 <- lm(Average_Rating ~ comfortable.with.XX.stars, data = player_avg_ratings)
model2 <- lm(Average_Rating ~ Current.Rank.SS.BL, data = player_avg_ratings)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```



	Model	R_Squared	P_Value
1	Comfortable Stars −−> Avg. Rating	0.007850329	0.6539032
2	Rank −−> Avg. Rating	0.008750683	0.6358802

Comfortable Stars vs. Average Rating
No significant correlation ($p = 0.444$, $R^2 = 0.0227$), indicating that players comfortable with higher star maps do not rate maps differently from those comfortable with lower stars. The regression line suggests a slight downward trend, but the low R^2 value indicates this effect is negligible.

Rank vs. Average Rating
No significant correlation ($p = 0.779$, $R^2 = 0.0031$), meaning that better-ranked players do not systematically rate maps higher or lower than lower-ranked players. The nearly flat regression line further confirms that rank has minimal influence on rating behavior.

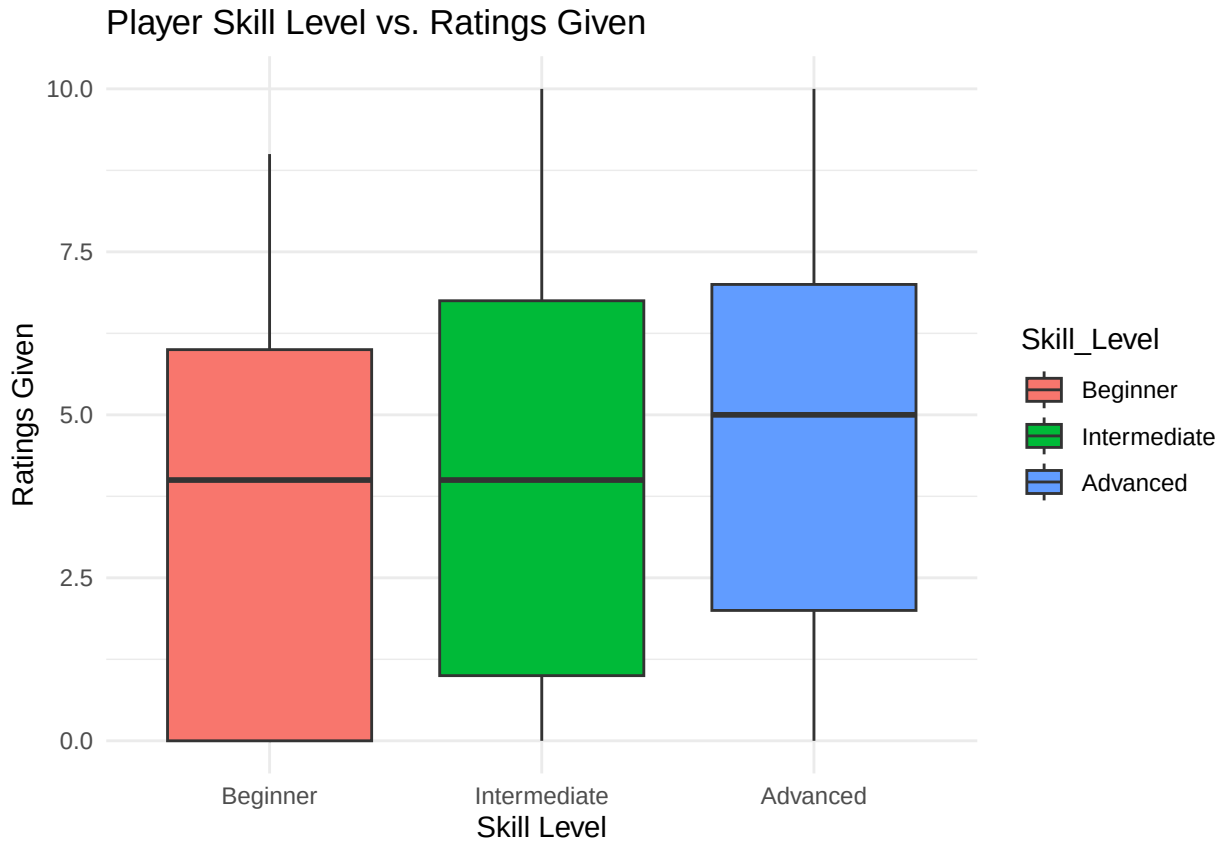
Conclusion

These results suggest that player preferences for map quality are **not directly linked** to skill level. Both high and low-skilled players rate maps similarly, implying that map enjoyment is subjective and not purely dependent on skill level or ranking experience.

Influence of Player Skill Level on Ratings Given

This boxplot visualizes how players of different skill levels (Beginner, Intermediate, Advanced) rate maps, based on the threshold **(-Inf, 7, 10, Inf)** for skill grouping. Unlike the scatterplots that examined individual correlations (Rank and Comfortable Stars vs. Ratings Given), this visualization categorizes players into skill groups, making it easier to compare overall rating distributions.

```
## Warning: Removed 14 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```



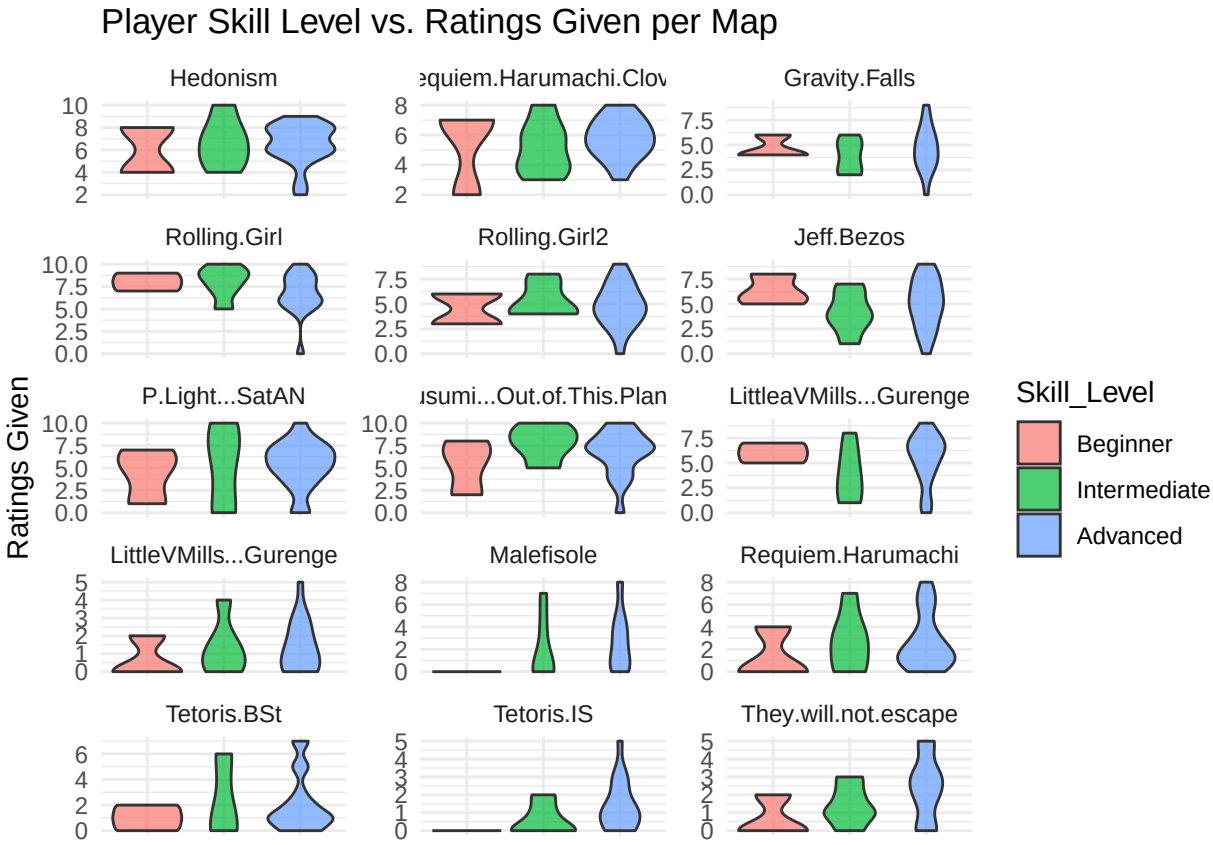
The previous models showed weak correlations (low R^2, high p-values), suggesting that rank or comfortable star level alone does not strongly predict rating behavior. But here the boxplot reveals broader trends:

- Beginners have a slightly higher median rating but also a large spread, indicating more varied opinions.
- Intermediate and Advanced players rate similarly on average, but the spread in Advanced players is larger, meaning they are more polarized in their opinions.

Beginners might rate higher on average because they are less critical or have lower expectations. Intermediate players show a balanced trend, suggesting that experience leads to a more neutral or measured evaluation approach. Meanwhile, advanced players may have more refined preferences, leading to more extreme ratings (both higher and lower).

Skill vs Rating per map

This analysis investigates how players of different skill levels (Beginner, Intermediate, Advanced) rate individual maps. The violin plots in the following Figure visualize the distribution of ratings across different skill groups for each map.



Skill level influences rating distribution, but not consistently across all maps. Some maps (`Rolling Girl, P\*Light - SatAN, Lusumi - Out of This Planet`) show distinct rating differences between skill levels, while others (`Gravity Falls, Jeff Bezos, Hedonism`) have similar ratings across groups. This suggests that certain maps divide opinions more strongly depending on skill level.

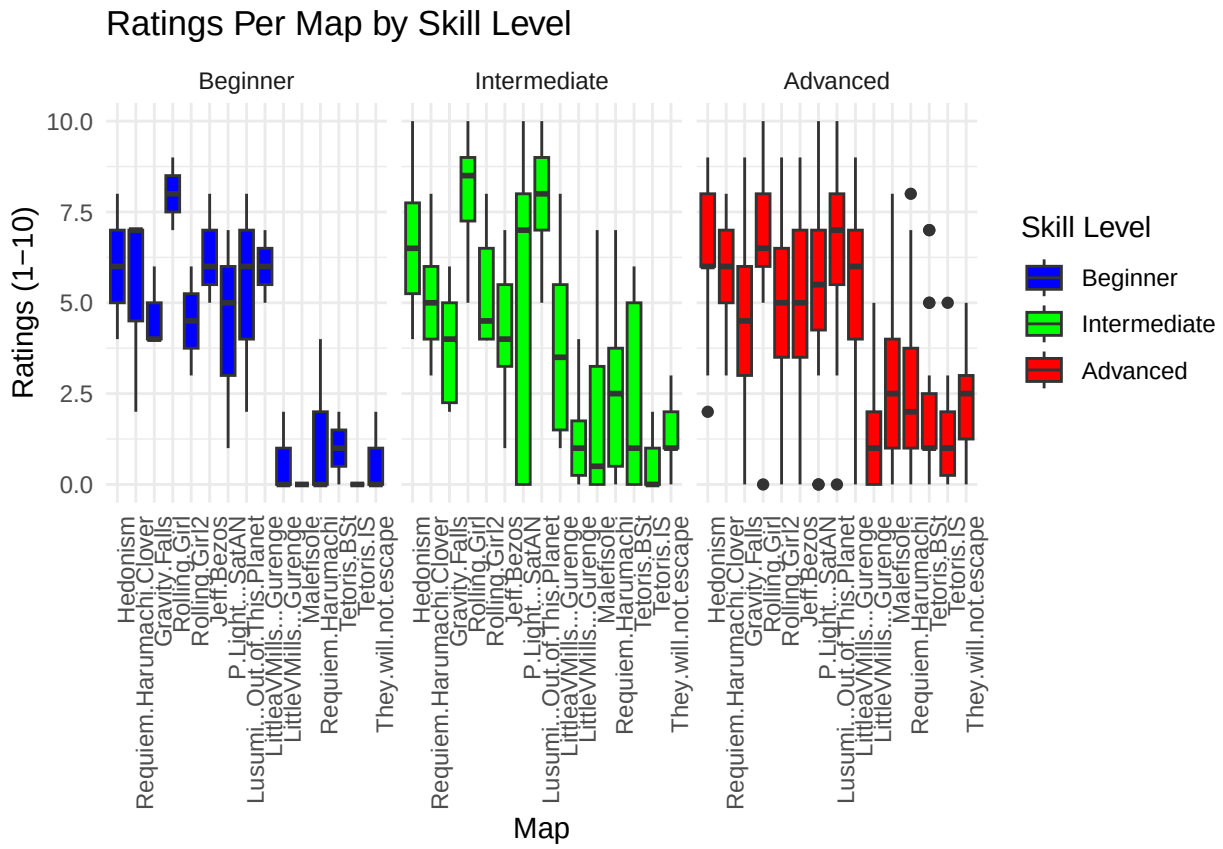
Tech and speed maps show higher variance among advanced players. Maps like Rolling `Girl, Lusumi - Out of This Planet, and P*Light - SatAN` exhibit larger rating spread among Advanced players, indicating that experienced players have more polarized opinions on complex maps. This could be due to a higher sensitivity to flow, mapping techniques, and difficulty.

Beginner ratings tend to be more uniform. Beginners often provide more centered ratings with less spread, particularly in maps like `Requiem Harumachi Clover and Jeff Bezos`. This suggests that less experienced players rate maps based on overall playability rather than nuanced mapping details.

AI-generated maps generally receive lower ratings, particularly from skilled players. `LittleVMills - Gurenge, Tetoris BSt, and They Will Not Escape` show low ratings from all skill groups, with Advanced players rating them the lowest. This aligns with prior findings that AI-generated maps tend to be rated worse, especially by experienced players who detect structural flaws more easily.

```
## `summarise()` has grouped output by 'Skill_Level'. You can override using the
## `.groups` argument.
```

Here is another Visualization of how the ratings differ:

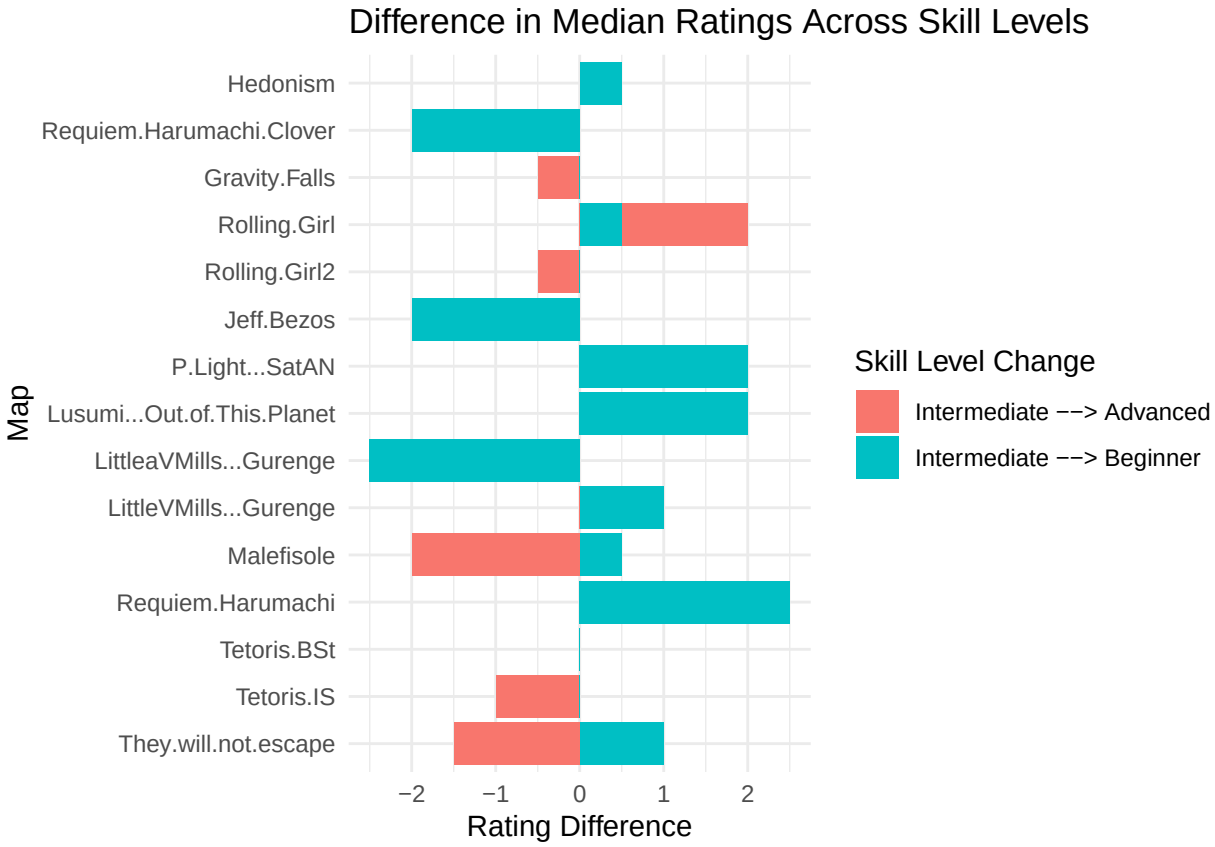


Conclusion:

- High-skill players tend to have a wider spread in ratings, particularly for technical maps.
- AI-generated maps receive lower ratings across all skill groups, but especially from Advanced players.
- Casual-friendly maps tend to be rated more consistently across skill levels.
- Future studies could explore whether specific mapping elements (e.g., flow, timing, resets) contribute to the observed rating variations.

Difference in Median Ratings Across Skill Levels

The following graph examines how median ratings change across different skill levels, comparing Intermediate to Beginner and Intermediate to Advanced players. The bar chart in the figure below visualizes these differences for each map.



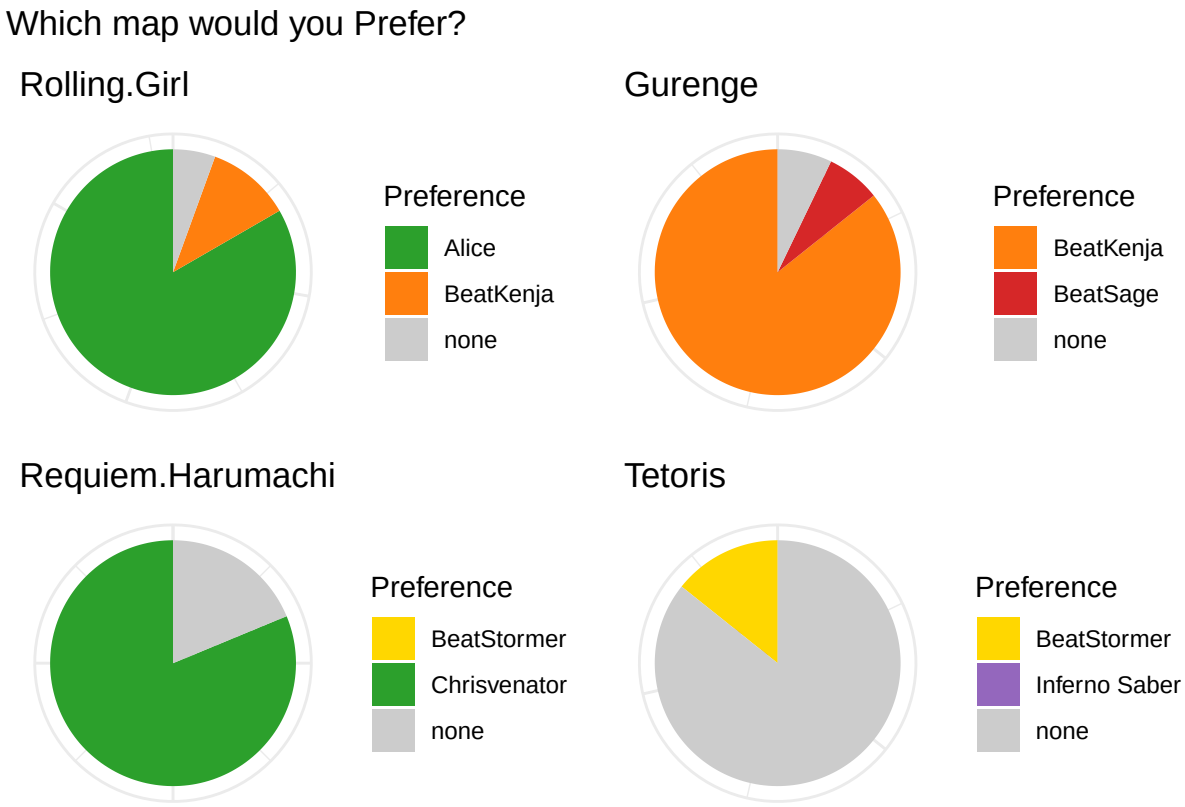
Advanced players tend to rate lower than Intermediate players, particularly on `They Will Not Escape, Tetoris IS, and Malefisole`, suggesting that more experienced players are more critical of certain maps. Beginners often rate higher than Intermediates, especially on `Lusumi - Out of This Planet and LittleVMills - Gurenge`, indicating that novice players may be less sensitive to mapping quality issues. Some maps (`Rolling Girl, Hedonism`) show minimal rating differences across skill levels, suggesting that certain maps are perceived consistently regardless of player experience.

Advanced players rate more critically, especially for AI-generated and complex maps. Beginners tend to give higher ratings, likely due to less exposure to high-level mapping quality. Some maps maintain stable ratings across all skill levels, suggesting universal (un-)playability.

These findings highlight that player experience influences rating behavior, particularly for AI-generated and highly technical maps even though there was almost no correlation in the R^2 and beta-values as shown before.

Comparison between the same map by different mappers

This analysis examines player preferences when choosing between different versions of the same map, created by different mappers/automappers. The pie charts in the figure represent the distribution of votes for each option, including cases where participants selected "none."



Human-made maps are preferred over AI-generated versions
`Rolling Girl (Alice)` is overwhelmingly preferred over `BeatKenja's` version, confirming that human mappers create better-received maps in terms of playability and structure. Similarly, `Requiem Harumachi` by Chrisvenator is strongly preferred over BeatStormer's AI-generated version, reinforcing that players recognize and appreciate human mapping quality.

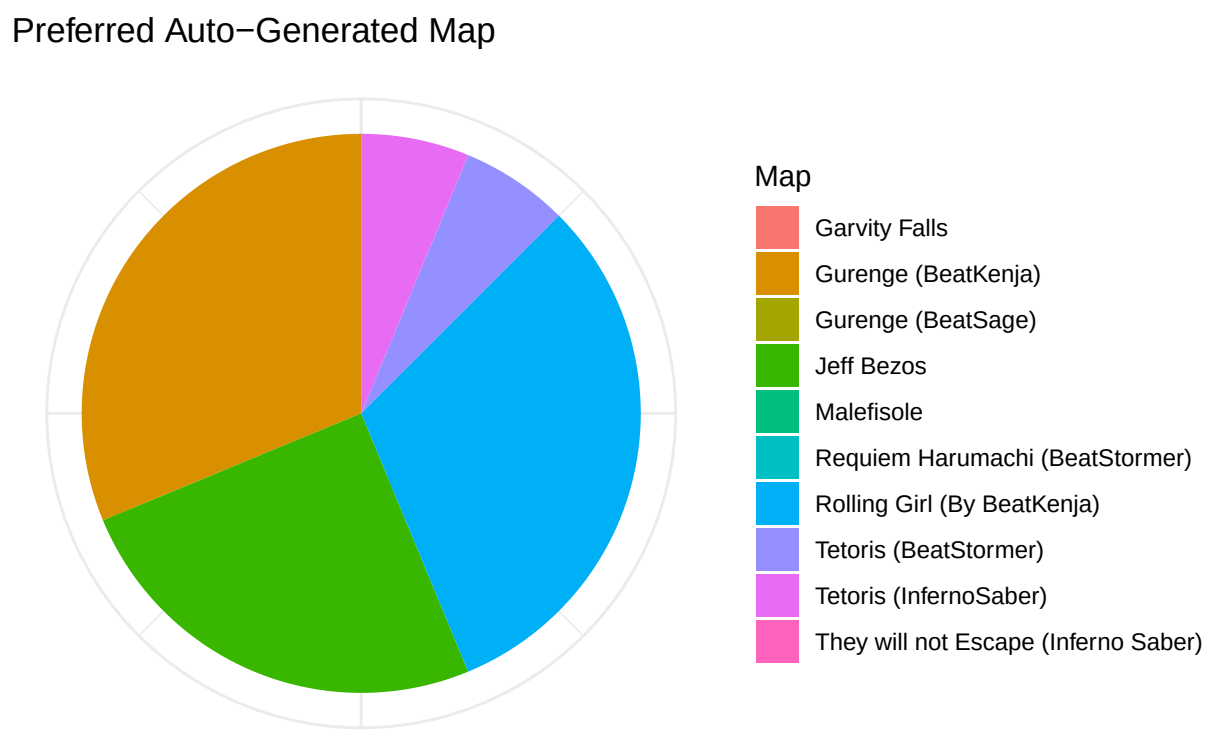
Players strongly favor BeatKenja over BeatSage
In the `Gurenge` comparison, BeatKenja receives nearly all votes, while BeatSage is barely chosen. This suggests that BeatSage's automapping has some features that make its maps less enjoyable. Further research is necessary as to why, like the interviews below.

A significant number of players abstained from choosing AI-generated maps
`Tetoris` (BeatStormer vs. Inferno Saber) shows that a large percentage of players preferred neither version, suggesting that both AI-generated maps had significant issues affecting playability. `Requiem Harumachi` (BeatStormer vs. Chrisvenator) also had a notable portion of "none" votes, but Chrisvenator's version was still favored.

Comparison to Previous Results The trends align closely with previous findings, where human-made maps consistently outperform AI-generated maps. BeatKenja, while an AI mapper, performs notably better than BeatSage, BeatStormer, and Inferno Saber, indicating that not all automappers perform equally. But it still loses against human-made maps. The presence of high "none" selections for AI maps suggests that, in some cases, players would rather not play than choose an AI-generated map.

Most liked map

In this question participants were asked which AI-generated map they enjoyed the most.



The pie chart titled "Preferred Auto-Generated Map" shows the distribution of preferences among auto-generated Beat Saber maps. `Gurenge (BeatKenja)` and `Rolling Girl (By BeatKenja)` dominate the preferences significantly, making up most of the pie. Maps such as `Jeff Bezos`, `Tetoris (InfernoSaber)`, and `Tetoris (BeatStormer)` have fewer preferences. The rest have no votes.

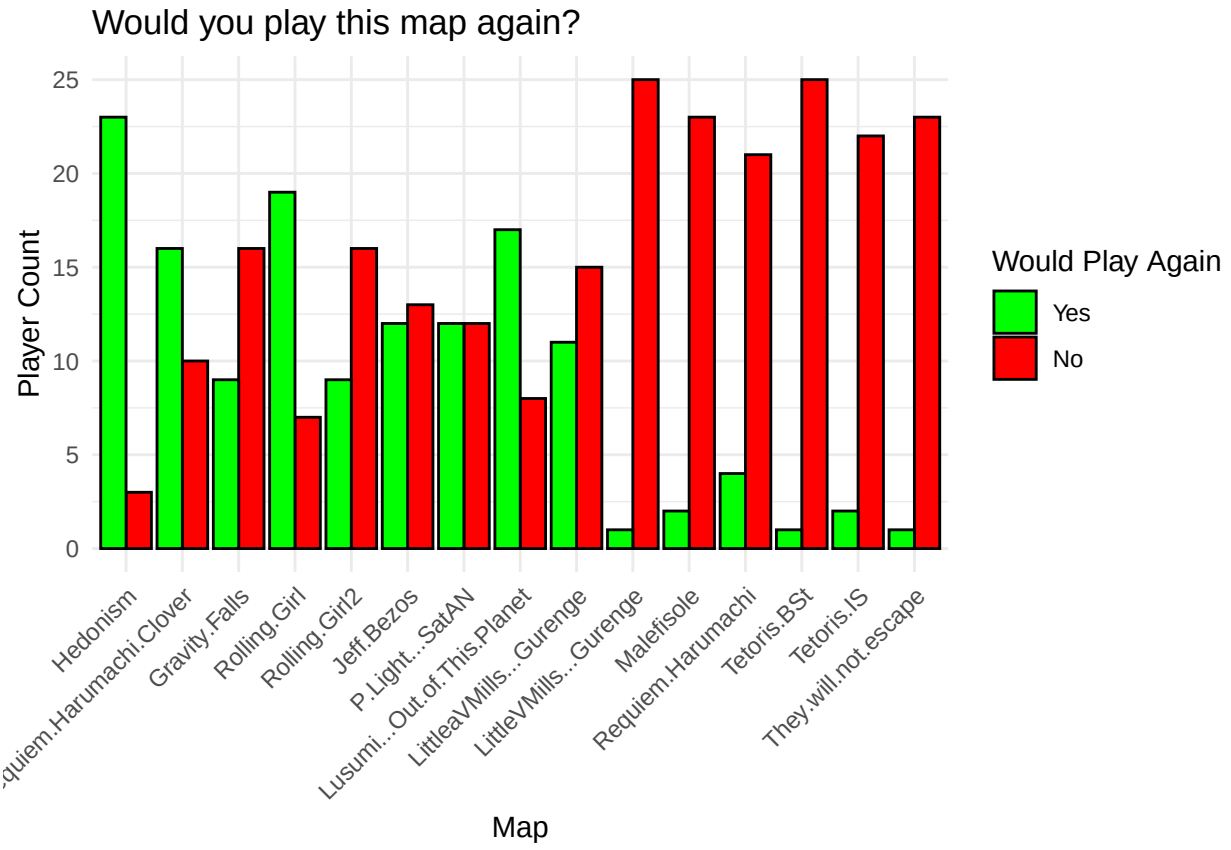
We also asked participants some questions and these were the outcomes:

- Gurenge (BeatKenja) is generally well-liked likely due to good timings and being an accuracy-oriented map.
- Rolling Girl (BeatKenja) is also popular, potentially because it combines timings from a well-liked human map, even though it has quirks such as speed and complex patterns.
- Jeff Bezos being slower and accuracy-focused, is less popular, possibly due to lower excitement.
- Inferno Saber maps (Tetoris.IS and They will not Escape) have fewer preferences, potentially due to awkward transitions and off-beat timings, making them uncomfortable to play.
- BeatStormer maps (like Malefisole and Tetoris.BST) might be less popular due to overmapping and unpredictable patterns.

Players highlighted timing issues, uncomfortable resets, vision blocks, and lack of flow as key negative factors for auto-generated maps. Participants clearly favored auto-generated maps that had accurate timings, better flow, and fewer unexpected or uncomfortable elements, resulting in BeatKenja's maps (`Rolling Girl` and `Gurenge`) being notably preferred.

Relation between rating and would play again

Participants have been asked if they would want to play a map a second time. The aim was to see, if the players would give a good rating even though they do not plan to play the map again.



Popular Maps

`Hedonism` (Human mapper, modified from BeatKenja's base): The most popular, with most participants willing to replay.

`Rolling Girl` (Human) and `Gurenge` (BeatKenja): These maps also have a favorable ratio of replayability, indicating strong player preference.

These three of them have no mapping errors. Even though `Rolling Girl` (Human) is a tech map, it still gets a good rating probably because it is tech is not too technical so even non-tech players are willing to play it.

Mixed Results

`Requiem Harumachi Clover`: Have moderate replayability. The player count wanting to replay is similar to or slightly higher than the count unwilling.

`Jeff Bezos` and `P*Light - SAtAN`: Almost equal distribution between willingness and unwillingness to replay, suggesting mixed feelings. A lot of participants told us they did not like `P*Light - SAtAN` because it is a speed map. It is very fast so skill may play a role. But further research is needed as to why.

Participants often answered that `LittleVMills - Gurenge (BeatKenja)` is quote "the most ok AI map I've ever seen". And participants generally liked it so it was a surprise to see them not wanting to replay the map. More research is needed as to why.

Unpopular Maps:

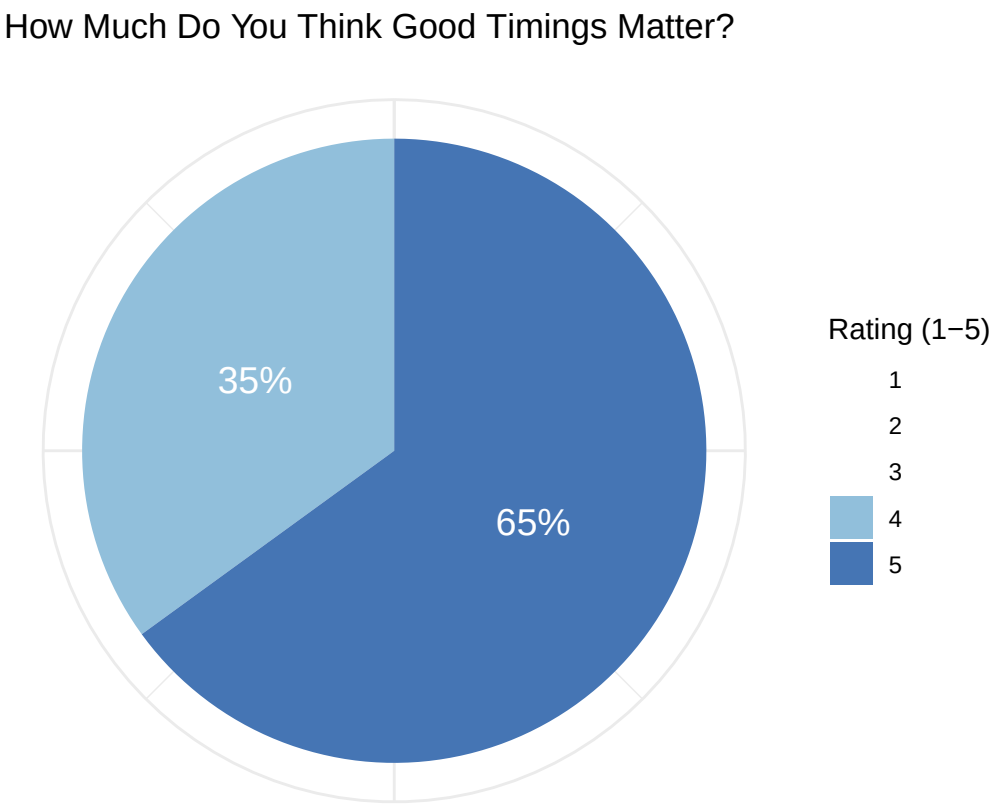
`Malefisole, Requiem Harumachi (BeatStormer), Tetoris (BeatStormer), Tetoris (InfernoSaber), They will not escape (InfernoSaber)`: These maps have overwhelmingly negative replayability scores, indicating players found them unappealing, possibly due to awkward patterns, bad flow, timing issues, or

overmapping. Participants often complained about having timing issues, being unpredictable, overmapping, unexpected resets, parity breaks, and generally bad flow.

Conclusion

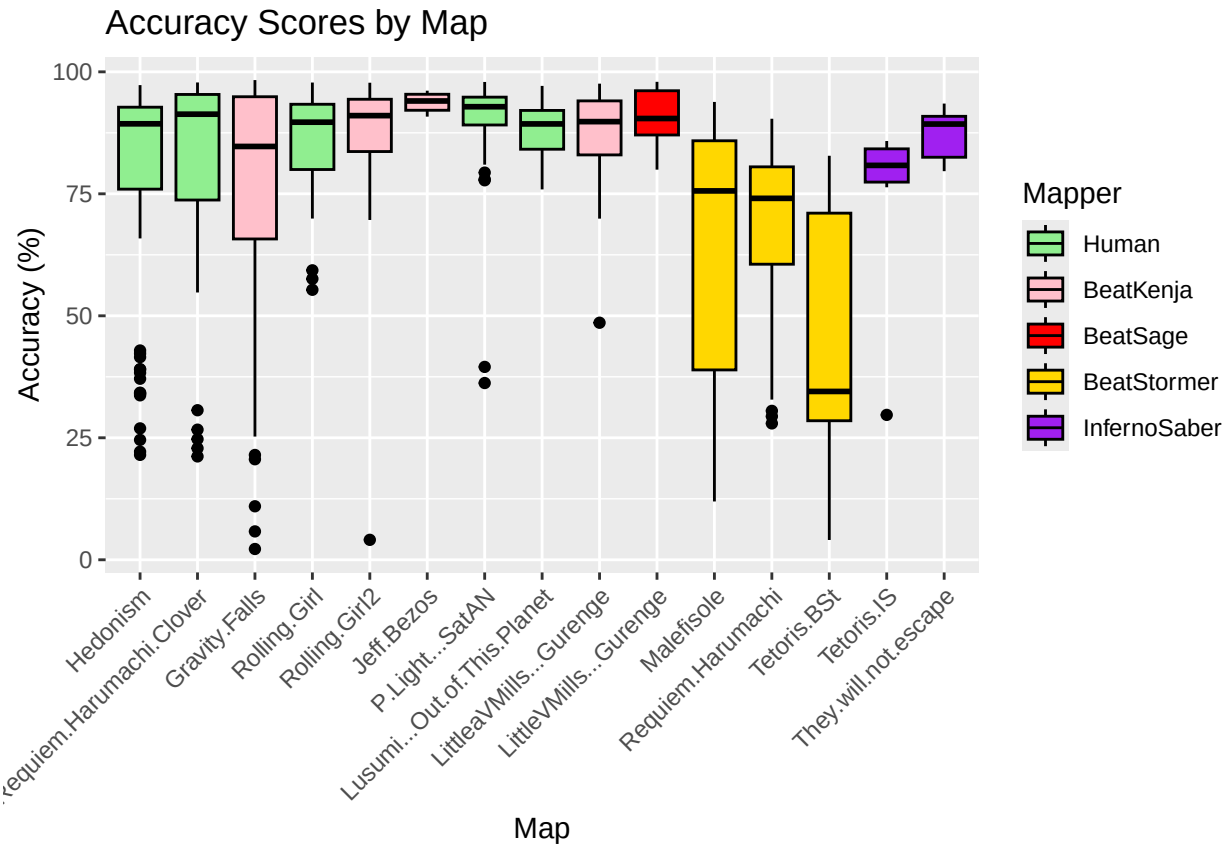
This data provides significant insights into map replayability, which is directly related to map quality. Maps with better timings, flow, and more intuitive or comfortable patterns (Hedonism, Rolling Girl, Gurenge (BeatKenja)) show high replayability.
Maps negatively rated tend to suffer from issues such as poor flow, uncomfortable transitions, overmapping, or awkward timings.
Replayability can act as an indirect measurement of the perceived fun and quality of Beat Saber maps, thus influencing the acceptance and popularity of auto-generated maps.

Do good timings matter?



Participants were asked, if they thought that good timings matter. The pie chart clearly illustrates the significance of accurate timings in Beat Saber maps according to player perception. 100% of participants rated the importance of good timings highly (either 4 or 5 out of 5), with 65% assigning the maximum importance (5) and 35% rating it 4. Thus, precise timing strongly influences players' assessment of map quality and enjoyment. This supports the idea that timing accuracy significantly impacts map quality and player satisfaction.

Scores per Map



The accuracy score distribution shows clear differences between mappers. Human-made maps consistently achieved high accuracy (85%–95%), suggesting intuitive flow and comfortable playability. BeatKenja maps displayed slightly lower medians and higher variability, reflecting occasional timing or pattern issues. Unexpectedly. the BeatSage map demonstrated moderate yet stable accuracy, despite known timing issues. In contrast, BeatStormer maps exhibited the lowest accuracy and high variability, indicating difficulty, overmapping, or uncomfortable transitions. InfernoSaber maps were intermediate, highlighting persistent patterns and timing shortcomings. Thus, mapper style significantly impacts player accuracy and map quality perception.

Note that there **may very likely be a survivership bias**! It is possible, that less experienced players might quit instead of submitting a score. Furthermore, players might also quit when the map is not to their liking. This results in only the scores of good players finishing the map and therefore skewing the graph visualisation upward.

Analysis of comments

Participants had the option to answer open-ended questions. This was purely optional. *ChatGPT was used to assist in identifying core themes, patterns, and participant insights. The findings were subsequently manually verified and rewritten.* The responses were grouped into *themes*. To preseve the anonymity of the players, they have been Assigned IDs and the IDs are randomized between each question.

Anything that bugged you about maps made by BeatKenja?

Several key themes emerged consistently from participant responses regarding BeatKenja's maps: **awkward or uncomfortable patterns**, **poor readability**, issues related to **song representation**, concerns about **pattern creativity**, and remarks on **flow and playability**.

The most frequently mentioned theme was related to the **awkward patterns**, appearing explicitly in the comments of participants 1, 2, 3, 5, and 11. Participant 1 noted "awkward parts" in maps like "Gravity Falls" and "Rolling Girl," emphasizing a significant decline in readability in later maps, describing them as "unreadable messes with resets and awkward patterning akin to a first map you would see in the testplays". It is important to note that the later maps this participant describes were not made by BeatKenja. Participant 2 echoed this sentiment, mentioning "patterns that were a bit wide or uncomfortable to hit," and Participant 3 provided a specific example of awkwardness through "patterns that were stacking into each other," creating confusion during gameplay. Participant 11 highlighted more technical issues such as "resets, upstarts, visionblocks," indicating difficulty in comfortably reading and executing the maps.

The second common theme was related to **poor musical representation**, specifically mentioned by participants 4, 5, and 12, with indirect references by participants 1 and 6. Participant 4 succinctly pointed out, "It needs better song representation," despite acknowledging the general mapping as "decent." Participant 12 echoed this critique by stating explicitly "not good musical representation." Similarly, participant 5 expressed concerns about timing and patterns ("choices in timing, triangles, JS/Offset"), indirectly referencing insufficient attention to musical cues. Thus, participants seemed to feel that the maps did not effectively match the music. Even though the general consensus was that the timings were onset, the patterns were not representing the music.

The theme of **pattern creativity** and overall quality also emerged strongly. Participant 12 explicitly mentioned "mostly no creative patterns," indicating dissatisfaction with perceived repetitiveness or lack of uniqueness. In contrast, participant 11 softened their critique by mentioning "some patterns were cool," suggesting occasional exceptions, but still reinforcing an overall deficiency in creativity. Participant 9 sarcastically remarked on linearity, stating "A tool can generate a map more linear than linear mappers do," indicating a perceived lack of originality in map design.

Concerns related to **flow and overall playability** were highlighted by Participant 10 who directly stated maps "didn't flow well imo," but acknowledged exceptions ("the Jeff Bezos map was fine"). Participant 1 further contrasts this by describing BeatKenja's maps as generally readable, while the other automappers became an "unreadable messes". This indicates that BeatKenja is not perfect but still better than every other Automapper that has been tested.

Lastly, it is important to mention that some participants provided neutral or positive comments. Participant 2, despite noting discomfort, concluded the maps were "a LOT better than BeatSage," indicating comparative superiority over another automapper. Participant 4 also minimized their complaint, stating their issue was "a small one," pointing toward generally acceptable quality despite their critique. Participant 18 had no negative feedback, simply stating "no."

In conclusion, the primary themes derived from participants' responses included awkwardness and discomfort in gameplay patterns, insufficient musical representation of the patterns, limited pattern creativity, and problematic flow. Despite these critiques, most participants acknowledged redeeming qualities, suggesting BeatKenja's automapper was considered acceptable overall but required improvements in several specific areas to meet player expectations consistently. Especially in contrast to the other three that were tested.

What do you think a map needs to be considered good?

The participants' responses highlighted several recurring themes regarding what makes a good map in Beat Saber: **good musical representation**, **flow and playability**, **creative and engaging patterns**, **consistency**, **technical correctness**, and to a lesser extent, **variety** and **subjectivity**.

The most frequently identified theme was **flow and playability**, explicitly emphasized by nearly all participants (participants 1, 3, 4, 6, 9, 10, 13, 15, and 16). Participant 4 succinctly stated that a map needs "good flow," while Participant 3 elaborated further, saying good flow occurs when the map has consistent patterns, such as a "gimmick (sideways inlines, etc.) that stays consistent." Participant 15 specifically connected flow to enjoyment, mentioning it requires "a form of continuity and fun." Similarly, participant 9 emphasized simplicity, clearly stating a map "needs good flow," reinforcing that smooth transitions between notes greatly influence player enjoyment.

Closely related to flow is the theme of **good musical representation**, which participants mentioned almost as often. Participants 1, 2, 4, 6, 7, 8, 13, 14, and 16 highlighted the importance of mapping style matching the music. Participant 1 provided a detailed explanation, emphasizing that a good map includes "good representation of the music through the mapping style (speed, tech, acc, challenge)." This sentiment was echoed by Participant 2, who explicitly pointed out the importance of alignment between map style and music genre, stating, "When you hear speed core it's not gonna be an acc map." Participant 7 reinforced that good maps must be "timed correctly" and that the representation of music makes maps memorable and replayable.

The theme of **creative and engaging patterns** emerged strongly across several responses (participants 4, 7, 8, 13, 15, 17, and 18). Participant 7 articulated this clearly, stating a good map should contain "fun and creative patterns that make you want to come back." Participant 17 succinctly emphasized that a good map includes "variety and imagination," underlining that originality in mapping enhances replay value and player enjoyment.

Technical correctness, another common theme, was specifically highlighted by participants 6, 8, and implicitly by participant 11. Participants repeatedly mentioned that a good map must be free from common errors such as "resets, upstarts, visionblocks" (participant 11), "no parity errors" (participant 8), and "no parity errors or other mapping mistakes." Technical correctness directly influences a map's readability and overall quality, as Participant 6 described: patterns should represent the music "in an interesting and fitting way" without errors interfering with gameplay.

Consistency was mentioned frequently (participants 1, 2, 3, 6, and 15). It is important that the whole map in itself is consistent and coherent. Whatevery gimmic a map has, it has to be consistent throught the whole map or else the map feels random. Participant 1 explicitly stated that patterns "make sense," and Participant 3 highlighted consistency as essential, particularly when special mapping elements or gimmicks are introduced.

Participants frequently acknowledged the inherently **subjective** nature of mapping preferences, explicitly noted by participants 1, 4, and 7. Participant 1 explained clearly, "In the end this question is biased; each player likes a different style," recognizing that opinions on map quality inherently vary. Participant 4 echoed this, admitting, "If a player likes it but I don't, it could still be a good map," highlighting that personal preference significantly impacts judgments of quality.

The theme of **variety** appeared somewhat less often but was notably highlighted by participants 2, 10, and 17. Participant 2 emphasized a dislike for repetitive or "copy-paste" mapping, advocating that variety makes a map better. Participant 17 similarly highlighted the importance of "variety," reinforcing that varied patterns are desirable.

Overall, these responses suggest that a "good" Beat Saber map, as perceived by players, primarily depends on strong musical alignment, smooth gameplay flow, creative but still consistent mapping, and **especially error-free patterns**. If this is achieved and the map is objectively good, individual preferences and subjective interpretation will play a significant role in determining if the player wants to play the map again.

In your opinion, what makes BeatSage maps bad?

The participant responses highlight several common themes regarding what makes BeatSage maps bad, with clear patterns emerging around issues of **poor flow**, **random and awkward note placements**, **lack of parity**, **timing inaccuracies**, and a general sentiment of overall dissatisfaction.

The most frequently cited theme was the issue of **poor flow**, explicitly mentioned by participants 5, 6, 8, 9, and 13. Participant 5 strongly criticized this aspect, stating "Everything, no flow," underscoring flow as a major weakness. Participant 8 similarly described BeatSage maps as having "bad flow," and Participant 13 directly stated "the lack of flow" as their primary complaint. This concern about flow connects closely with another prominent theme: **randomness and awkwardness**.

The issue of **random and awkward note placements** appeared frequently, explicitly mentioned by participants 3, 8, 9, 11, and 14. Participant 3 described maps as having "jumbled notes," emphasizing their unreadability and noting specifically "resets and just jumbled notes" that disrupt gameplay. Participant 8 reinforced this critique, remarking that BeatSage maps have "random angles," leading to awkward patterns that don't align logically with the gameplay experience. Participant 14 succinctly stated "random patterns," further reinforcing the concern of unpredictability as a negative trait of BeatSage-generated maps.

Participants 11 and 12 explicitly raised **parity issues**, with Participant 11 mentioning "lack of parity" directly, and Participant 12 calling out "terrible parity," making maps "barely playable." This aligns closely with concerns of awkwardness, as incorrect parity disrupts player experience by causing unexpected or unnatural movements and may even cause injury.

Participants also frequently noted **timing inaccuracies** as a negative trait. Participant 4 specifically critiqued BeatSage's tendency toward a "stagnant 120bpm being auto-chosen no matter what," which negatively affects timing accuracy and thus the musical representation. Participant 9 added to this criticism, noting "some mistimings," directly connecting timing problems with diminished playability.

Participants 1, 5, and 15 expressed particularly strong overall dissatisfaction, criticizing virtually every aspect of the BeatSage maps. Participant 1 humorously summarized their frustration, stating, "Everything xD... They're the worst in all regards," expressing a strong general dissatisfaction. Participant 16 echoed this by humorously adding "everything lol," suggesting a general negative perception, although less explicitly detailed. This broad negativity highlights an overarching sense of disappointment, often due to multiple issues compounding in a single map.

Interestingly, some participants did not entirely dismiss BeatSage maps, offering qualified criticism instead of complete rejection. Participant 4 pointed out that "not ALL of them are bad," despite voicing a specific frustration. Similarly, participant 8, while criticizing "bad flow" and randomness, did not entirely write off the maps, implicitly suggesting occasional redeeming features despite significant flaws.

In summary, the participants consistently identified poor flow, randomness, lack of parity, timing inaccuracies, and a general sense of dissatisfaction as key shortcomings of BeatSage maps. Although responses varied in intensity, it was clear that the randomness and lack of thoughtful patterning significantly detracted from the perceived quality and playability of these maps.

What are mistakes that auto-mappers make that make a map unplayable/not enjoyable?

The participant responses from the user study clearly highlight several recurring themes regarding negative aspects of BeatSage-generated maps in Beat Saber. The main issues participants emphasized were **resets and uncomfortable patterns**, **timing inaccuracies**, **mapping inconsistency (over- and undermapping)**, **randomness and poor note placement**, and problems with **parity and readability**.

The most prominently recurring theme was the presence of **resets**, which are unexpected changes in the direction of notes that disrupt natural gameplay flow. Participants explicitly mentioned resets frequently, notably participants 3, 9, 15, 16, 17, and 18. For instance, participant 3 listed multiple specific issues including "resets, vision blocks, stacked notes," explicitly categorizing resets as a core issue alongside other mapping mistakes. Participants 15, 17, and 18 mentioned resets as the primary issue without further elaboration, emphasizing their widespread and disruptive nature ("resets, mostly" [17]; "resets" [18]). Participant 1 implicitly connected resets with timing discrepancies and inconsistency, especially noticeable in faster, bass-heavy songs, stating "there seems to be a discrepancy of representation" that contributes to readability issues.

Closely linked to resets were **awkward and uncomfortable patterns**, explicitly mentioned by participants 1, 4, 9, 10, and 16. Participant 16 specifically pointed out "overly wide patterns," while participant 9 identified both "wide patterns" and "other notes in the cut-path," creating uncomfortable movements. Participant 2 also alluded to this indirectly through mentions of uncomfortable hits in other contexts, supporting this theme as a significant issue.

Timing inaccuracies and **Wrong music representation** formed another important concern, clearly addressed by participants 1, 3, 8, 9, and 10. Participant 1 specifically highlighted that for "higher noise maps, faster or bass boosted" tracks, there was a significant "discrepancy of representation," creating difficulty identifying notes relative to the music. Participant 3 strongly expressed similar concerns, criticizing maps as being "clearly off-time" and mentioning problematic aspects like "overmapping slower sections, undermapping faster sections." Participant 6 directly criticized mistimed notes, describing "clearly off-time parts," while participant 10 complained about patterns that did "not fit the beat."

Another strongly expressed concern was **randomness and poor note placement**, explicitly mentioned by participants 4, 8, 10, 11, 13, and 14. Participant 10 described maps as having "just random notes," participant 4 mentioned the introduction of unnecessary difficulty through "unnecessary diagonal or sideways angles," and participant 11 described it succinctly as "random note placements," reflecting how randomness disrupts logical gameplay experience. Participant 8 similarly complained about "blocks on the offbeat" inconsistently.

The issue of **poor parity and readability**, involving uncomfortable or unclear alternations between hands, emerged strongly. Participant 11 explicitly described maps as "2018 style of no parity," emphasizing a regression to outdated mapping practices. Participant 12 reinforced this theme, clearly highlighting "terrible parity, maps barely playable," suggesting the seriousness of parity problems in affecting gameplay. Participant 8 further supported this by noting "unfitting parity breaks," and participant 9 pointed to resets and parity breaks indirectly through concerns about the overall readability ("notes in the cut-path").

Notably, there is some variation in participant criticism: participants 1 and 4 nuanced their critiques by indicating circumstances under which BeatSage maps could still be somewhat acceptable ("easier maps or ones where the automapper doesn't have errors," participant 1; "not ALL of them are bad," participant 4). This contrasts sharply with the harsher criticism expressed by participants like 5 ("Everything, no flow") or participant 12, who described maps as having an overall "outdated" and unplayable character.

In summary, participants primarily identified **resets**, **awkward and uncomfortable note patterns**, **poor timing**, **randomness**, and **poor parity** as significant flaws that detract heavily from the playability and enjoyment of Autgenerated maps.

In your opinion, what should auto-mappers do so that the generated maps improve in quality?

The participant responses regarding potential improvements for automappers like BeatSage reveal clear recurring themes, prominently including **parity improvements**, **pattern consistency and coherence**, **better musical representation**, **flow enhancements**, **injury prevention**, and suggestions about **manual editing or training with current mapping standards**.

The most emphasized theme was related to **parity improvements**, explicitly mentioned by participants 1, 8, and 12. Participant 1 strongly advocated for "Getting rid of triangles and DDs," explaining that this change would greatly improve player experience. Participant 12 succinctly reinforced this, suggesting automappers need to "learn parity," highlighting parity's importance to mapping quality. Participant 8 provided a health-oriented perspective, emphasizing removing certain parity-breaking patterns, as they "can injure or cause pain," thus tying parity directly to player safety and comfort. These players also stated that maps by BeatKenja were "not bad parity wise".

Another major theme was the need for **pattern consistency, coherence, and flow improvements**, specifically addressed by participants 1, 3, 5, 9, and implicitly by others. Participant 1 pointed out inconsistencies, noting that automappers should "ease up on random difficulty spikes," addressing abrupt difficulty changes without the song accompanying it. Participant 5 recommended developing "a catalog of commonly used patterns" to prevent random single-note placements that lack correlation, thus improving coherence in the map itself. Similarly, Participant 9 mentioned "consistency and patterns that make sense," suggesting that coherent and predictable patterns are vital. Participant 4, referencing mapping trends, stressed that automappers should "Keep their database up to date," implying that consistency requires adaptation to current mapping standards rather than outdated styles.

Closely linked to consistency is the theme of improving **musical representation and emphasis**, frequently highlighted by participants 1, 10, 11, and 12. Participant 10 explicitly advocated manual revisions and manual cleaning to ensure "proper emphasis," indicating a strong need for maps to align better with musical highlights. Participant 11 similarly stated that good maps have "interesting patterns and good flow that matches with the music," clearly linking creativity, flow, and musical representation together. Participant 13 also suggested improving the relationship between notes and music, advocating simply, "Make a better song representation."

A related theme was the necessity for automappers to maintain **up-to-date mapping standards**, directly stated by Participant 4, who recommended automappers "keep their database (?) up to date because mapping standards change." They suggested maps felt outdated ("from 4-5 years ago"), emphasizing the evolving nature of mapping practices and player expectations.

Participant 3 introduced a practical suggestion by recommending **technical accuracy** and specifically addressing "Incorrect bpm timing," urging automappers to avoid timing mistakes by accurately extracting song data. They pointed out the practical issue that "high noise maps will confuse" the automapper, highlighting a technical improvement area. This concern was also indirectly supported by Participant 1, who described the timing as inconsistent and misrepresenting the music, implying the need for more sophisticated audio analysis methods.

Lastly, participants identified a notable distinction between automapped and human-mapped maps, highlighting the issue of perceived randomness or "lack of soul," implicitly suggested by Participant 1 ("feels like someone gave AI timings without actually hearing the song") and Participant 14, who emphasized random patterns negatively. This highlights an underlying theme that players desire a more intentional, purposeful feel to mapping, often perceived as lacking in automated processes.

Overall, the responses indicate clear pathways to improvement: automappers would significantly benefit from addressing parity errors, developing coherent and consistent patterning, improving musical representation through accurate timing and appropriate emphasis, ensuring patterns align with contemporary mapping standards, and potentially incorporating selective manual refinements.

How would you describe good flow? How do you think flow can be emulated?

Participants described **good flow** predominantly through several closely connected themes: **natural transitions between notes**, **consistency and predictability**, **effective representation of music**, **momentum in swings**, and avoiding **awkward disruptions or resets**.

The most consistently emphasized theme was **natural and intuitive transitions between notes**, explicitly highlighted by participants 1, 2, and 8. Participant 2 provided a clear description, stating flow is present "if the player can hit the next note easily from where the last note ended to where the next note's swing starts," emphasizing seamless movement between notes. Participant 1 similarly highlighted the importance of preparation for upcoming notes, stating good flow involves a "good setup into the next note," proposing a practical method for automappers: "connecting blocks with lines to detect and correct overly sharp turns" (Participant 1).

Closely related was the theme of **avoiding awkward disruptions or resets**, explicitly mentioned by participants 4, 5, 7, and 10. Participant 10 succinctly described good flow as having "no random map style changes and no RESETS," reinforcing the idea that abrupt changes disrupt flow significantly. Participant 5 similarly described good flow as a map that "doesn't surprise you with something you didn't expect AT ALL," stressing predictability. Participant 4 emphasized comfort and continuity, describing flow as "when a map plays itself," without needing constant intense focus due to sudden, unexpected changes.

Participants frequently connected flow explicitly to the **effective representation of music**, notably highlighted by Participants 3, 6, 8, 9, and 11. Participant 9 provided a clear emotional description: "Good flow is when you feel like you are playing the song," or like "You are playing the song to a certain degree," directly linking flow to musical immersion. Similarly, Participant 11 simplified this idea as "patterns that match the music," implying musical alignment inherently contributes to a sense of good flow.

The concept of **momentum and rhythmic alignment** emerged explicitly, with Participant 8 emphasizing that "a lot of good flow has to do with the swing momentum AND the timing of the notes combined," reinforcing that flow involves both physical motion and precise rhythmic accuracy. Participant 2 echoed this by emphasizing the ease of transitioning swings from one note to the next, highlighting the necessity of mapping notes in physically comfortable sequences.

The issue of **awkward disruptions or resets** was clearly and frequently highlighted as a major detractor from flow (Participants 5, 10). Participant 10 succinctly defined good flow as the absence of "random map style changes and no RESETS," suggesting resets represent a critical obstacle disrupting natural play. Similarly, Participant 5 described disruptions as "breaking the whole vibe of a map".

Interestingly, while participants focused primarily on defining good flow, some provided practical advice or ideas for emulating flow in automappers. Participant 1 suggested an algorithmic approach, recommending measuring and limiting the curvature of note placements. Participant 2 defined good flow as the ability to "hit the next note easily from where the last note ended," providing an intuitive and simple heuristic automappers might implement.

In summary, participants characterized good flow through natural transitions, consistency, predictability, rhythmic alignment, momentum, and musical coherence. These elements combine to create an intuitive, enjoyable experience, fundamentally opposing sudden disruptions, resets, or unpredictable note placements. The participants provided both abstract definitions and practical suggestions.

Other comments

The most prominent theme was the perceived **lack of creativity and human intentionality** in AI-generated maps compared to those crafted by human mappers. Participant 1 emphasized that AI-generated maps felt "hollow" and devoid of "creativity and imagination," contrasting these negatively with human-made maps, which were characterized by "intentions and directions" that allowed players to "imagine" and experience the game with greater engagement. This sentiment was echoed by Participant 2, who remarked that AI-generated maps felt as if they were produced without genuine enthusiasm or enjoyment, stating, "It's like someone was forced to map something that they don't like."

Another significant theme centered on the perception of **playability and enjoyment**. Several participants expressed dissatisfaction with the overall quality and playability of the automapped content. Participant 5 acknowledged improvement in automapping compared to the past but still found AI-generated maps "barely playable." More explicitly negative feedback came from Participant 6, who described the maps from BeatSage, BeatStormer, and InfernoSaber as "atrocious," reflecting strong dissatisfaction with their experience.

Interestingly, the theme of **hybrid mapping** (combining AI and human effort) emerged positively. Participant 3 specifically praised the hybrid approach used in the map "Gurenge" by BeatKenja, emphasizing its superior quality and enjoyable gameplay experience, labeling it as their "favorite autogenerated/hybrid map" to this day. This indicates a preference for hybrid maps, suggesting that when human oversight complements AI mapping, the results significantly improve.

BeatKenja's maps in particular **received notable positive attention**, with Participant 7 summarizing their impression as "overall solid maps by BeatKenja," reinforcing Participant 3's praise. The specific positive mention of BeatKenja four times among seven participants who answered this question indicates a clear appreciation for the mapper's hybrid approach.

In **summary**, the analysis highlights three central themes: the perceived absence of creativity and intentionality in purely AI-generated maps, significant concerns regarding playability and quality of AI-generated content, and positive feedback toward hybrid maps—particularly those created or refined by specific mappers like BeatKenja. The frequency of comments strongly supports these themes, with the majority of participants (five out of seven) criticizing AI-generated maps for lacking human elements and being less enjoyable, while hybrid approaches received consistently higher praise.

Prompt used

This Prompt was used to help summarize and better formulate the previous interview answers mentioned. The model used was ChatGPT4.5.

```
I conducted a user study for my thesis, where participants answered the following
open-ended question: '<Question>'. The topic was the comparison between different
automappers for the Rythm Game "Beat Saber".

Below are the participant responses:
<Responses>

Please provide a detailed qualitative analysis of these responses in a continuous
text narrative format. Clearly identify and explain the main themes that emerged,
describe how frequently each theme appeared, and discuss any relevant connections
or notable contrasts among the themes.

Within your analysis, please integrate direct quotes from participants to reinforce
your statements. Clearly cite each quote using the corresponding Participant ID.
```

Interview Conclusion

The analysis revealed that participants frequently highlighted BeatKenja's automapper positively, especially praising its superior overall quality compared to other automappers. Despite identifying specific issues like awkward patterns, insufficient musical representation, limited creativity, and problematic flow, participants generally acknowledged that BeatKenja produced maps with better parity and fewer gameplay disruptions, making them more comfortable and readable. Several players appreciated the hybrid mapping approach used in BeatKenja maps, particularly emphasizing the enjoyable patterns and good musical representation found in examples like "Gurenge."

Good maps were consistently described as those with excellent musical representation, smooth flow, creative and engaging patterns, technical correctness, and consistency.

In contrast, BeatSage, BeatStormer, and InfernoSaber maps were often criticized for poor flow, randomness, awkward note placements, timing inaccuracies, and parity issues, leading to widespread dissatisfaction.

Suggestions for improving automappers emphasized the necessity of addressing parity errors, ensuring pattern consistency, enhancing musical accuracy, updating mapping standards, and reducing awkward note placements.

Finally, participants strongly preferred hybrid mapping approaches combining AI with human input, praising these for significantly higher creativity, intentionality, and overall enjoyment compared to purely AI-generated maps.

Conclusion

This study provides an in-depth analysis of auto-generated and human-made Beat Saber maps, incorporating both quantitative rating data and qualitative feedback from participants. Our findings highlight key differences between mapping styles, player preferences, and the overall reception of AI-generated content compared to manually created maps.

Key Findings

Human-made maps consistently received higher ratings and had lower abort rates, indicating superior playability, better flow, and stronger musical representation. In contrast, AI-generated maps, particularly those from BeatStormer, InfernoSaber, and BeatSage, were rated lower and exhibited a higher frequency of early exits. The most common issues cited for AI-generated maps included awkward patterns, inconsistent flow, mistimed notes, and poor parity.

BeatKenja's automaps showed mixed results. Some maps, such as `Gravity Falls` and `Rolling Girl`, received polarized ratings, suggesting that they were enjoyable to some players but unappealing to others. Despite significantly outperforming the other AI mappers, BeatKenja's maps still fell short when compared to their human-made counterparts.

Player Skill and Rating Trends

Interestingly, there was no significant correlation between player skill level and rating behavior. Both high and low-ranked players exhibited similar rating distributions, reinforcing the idea that enjoyment of a map is highly subjective. However, advanced players tended to rate tech maps more critically, while beginners provided more uniform ratings across all map types. The reason my be that advanced players have already preferences like for example speed, tech or acc.

Player Preferences and Future Directions

Participants identified key qualities that define a "good" map: smooth transitions, consistent patterning, and accurate musical representation. AI-generated maps frequently lacked these attributes, making them less favorable compared to human-designed maps.

A particularly notable insight was the preference for maps with strong flow and intuitive note placements. Players valued maps that felt natural and were structured logically in alignment with the music. AI-generated maps, especially from BeatSage, often failed in this regard, leading to frustration and higher dropout rates.

Implications for Automapper Development

The study's findings suggest several areas for improvement in automapper design. To create more enjoyable AI-generated maps, automappers should prioritize:

- **Better musical representation**: Ensure that note placements align accurately with beats and rhythms.
- **Pattern consistency**: Avoid unnecessary difficulty spikes and maintain logical note structures.
- **Parity improvements**: Reduce awkward resets and avoid parity-breaking patterns.
- **Enhanced flow**: Optimize note placements to create smoother and more intuitive transitions between movements.
- **Updated mapping standards**: Incorporate contemporary best practices used by experienced human mappers.

Final Thoughts

While current automappers can produce playable maps, they still fall short of human quality in critical areas. BeatKenja demonstrated notable improvements over other AI mappers, yet there remains room for further refinement. Future advancements in AI-driven mapping, potentially through machine learning or reinforcement learning, could enhance the quality of auto-generated maps to rival those made by human mappers.

Ultimately, the study underscores the importance of human intuition and expertise in map design. Until AI can effectively replicate the nuanced decision-making of human mappers, manually crafted maps will likely continue to provide the best player experience. Future research could explore hybrid approaches where AI-generated skeletons are refined by human intervention to bridge the gap between automation and high-quality mapping.