

W205 Final Project: Cricket Fielding Statistics

Chula Watugala

Introduction

Description

Cricket is a sport rich with statistics – probably one of the most numbers-focused games around. However, there is a glaring lack of fielding statistics in the game – something that a similar sport like Baseball has handled by the use of video analysis after the game and error statistics.

No such setup exists in cricket – mostly because there is the perception that fielding is too subjective to evaluate. Compared to batting - where it's about getting runs and you can get a simple batting average measure on how many runs you get every time you bat, or bowling - where you can measure how many runs you give away per out, for fielding there's no clear metric to evaluate players. Because of this there are a lot of myths in the game when it comes to fielding - commentators say "he would've caught that 99 times out of a 100" when some fielder drops a catch. Even though it's just an expression, there is no data on what the actual drop rates of fielders are - or what is an actual acceptable drop rate.

Objectives

The objective of this project is to generate cricket fielding statistics using live text commentary feeds. Fielding events and the players involved will be identified and parsed from the text and the data will be used to create useful fielding statistics. If this allows for a method to value fielding performance accurately, franchises would be able to evaluate player value more accurately.

Fielding events considered:

- Dropped Catches
- Misfields
- Missed Stumpings
- Direct Hits (Run-outs)
- Great Catches
- Catches
- Runs Saved

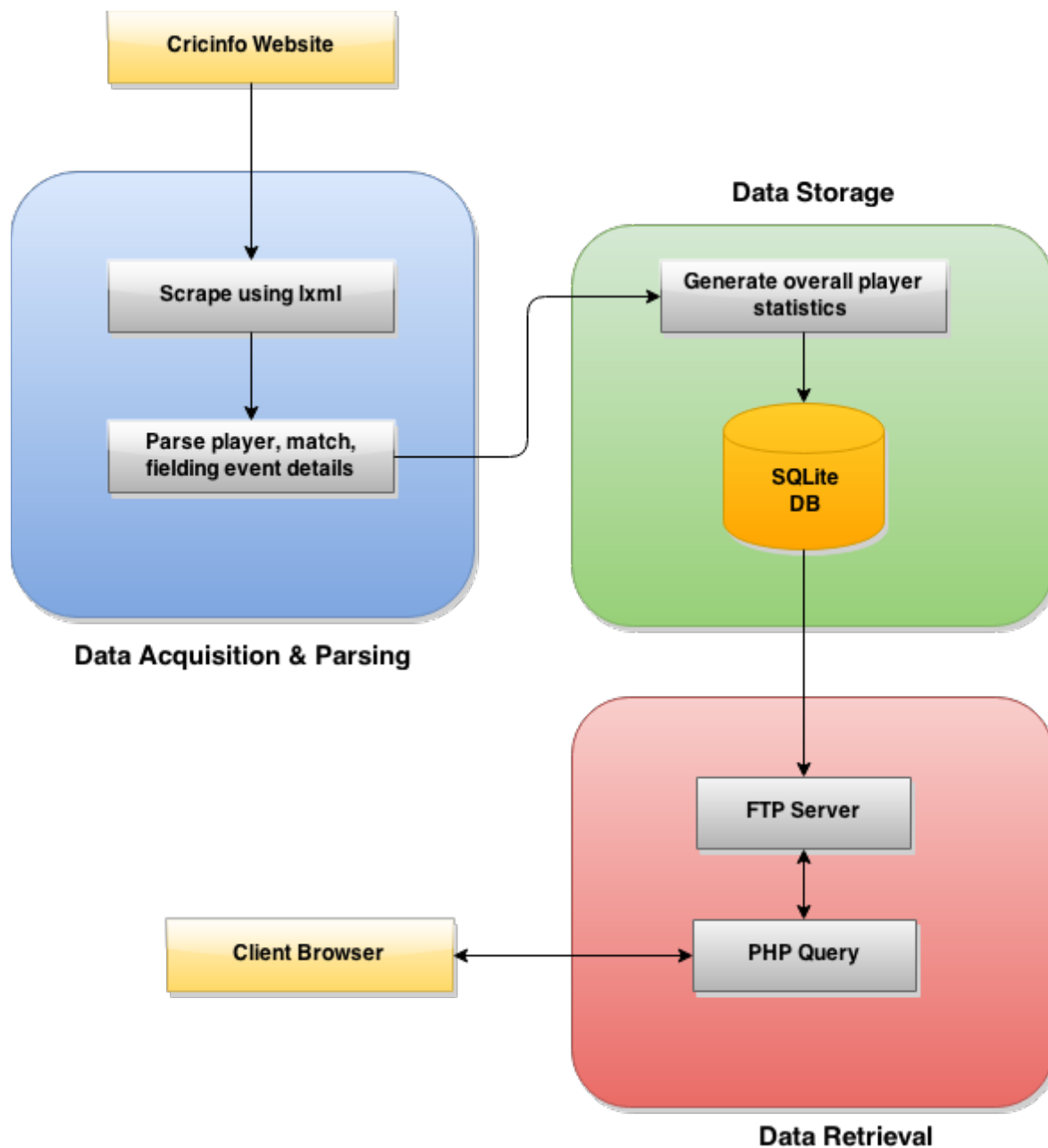
Data Sources

There are multiple sites like [Cricinfo](#), [Cricbuzz](#) and [BBC Cricket](#) that offer live text commentary during games, but only Cricinfo has done it for a long period of time (10+ years) and has had a consistent structure that enables systematic parsing.

Related Work

There have been multiple calls for the necessity of fielding statistics ([here](#), [here](#) and [here](#)), but no concerted effort has been made to do this using an automated and systematic method till now. Therefore this project has the potential to revolutionize how cricket values the fielding aspect of the game.

Solution Architecture



- *Data Acquisition & Parsing:* The lxml package in python is used to scrape each historical match with text commentary and parse the data in xml form.
- *Data Storage:* The generated data is stored in a SQLite database which python has built-in support for.
- *Data Retrieval:* Once the data is generated, it will be hosted on a website using PHP to query the FTP server.

Data Pipelines

The match results list has links to each match scorecard directly on the page. These links are scraped for match details and stored in a *MatchInfo* table dedicated for match details with a unique ID with the teams involved.

Match results						
Team 1	Team 2	Winner	Margin	Ground	Match Date	Scorecard
Afghanistan	Scotland	Afghanistan	8 wickets	Dubai (CA)	Jan 8, 2015	ODI # 3572
Afghanistan	Ireland	Ireland	3 wickets	Dubai (DSC)	Jan 10, 2015	ODI # 3573
New Zealand	Sri Lanka	New Zealand	3 wickets	Christchurch	Jan 11, 2015	ODI # 3574
Ireland	Scotland	Ireland	3 wickets	Dubai (DSC)	Jan 12, 2015	ODI # 3575
Afghanistan	Scotland	Scotland	150 runs	Abu Dhabi	Jan 14, 2015	ODI # 3576
New Zealand	Sri Lanka	Sri Lanka	6 wickets	Hamilton	Jan 15, 2015	ODI # 3577
Australia	England	Australia	3 wickets	Sydney	Jan 16, 2015	ODI # 3578
South Africa	West Indies	South Africa	61 runs	Durban	Jan 16, 2015	ODI # 3579
New Zealand	Sri Lanka	no result		Auckland	Jan 17, 2015	ODI # 3580
Afghanistan	Ireland	Afghanistan	71 runs	Dubai (DSC)	Jan 17, 2015	ODI # 3581
Australia	India	Australia	4 wickets	Melbourne	Jan 18, 2015	ODI # 3582

Match List



MatchInfo
Id
Team1
Team2


Each match scorecard looks like the example shown below where you get a list of the players involved and their performances with bat and ball. Each player name is linked to a profile page with his unique details, so these player names are scraped to populate a unique player information table.

ICC Cricket World Cup - 42nd match, Pool B
Ireland v Pakistan

ODI no. 3639 | 2014/15 season
Played at Adelaide Oval (neutral venue)
Pakistan won by 7 wickets (with 23 balls remaining)
15 March 2015 - day/night match (50-over match)

Ireland innings (50 overs maximum)		R	M	B	4s	6s	SR
WTS Porterfield*	c Shahid Afridi b Sohail Khan	107	161	131	11	1	81.67
PR Stirling	lbw b Ehsan Adil	3	15	8	0	0	37.50
EC Joyce	c Umar Akmal b Wahab Riaz	11	41	18	1	0	61.11
NJ O'Brien	c Umar Akmal b Rahat Ali	12	20	10	2	0	120.00
A Balbirnie	c Shahid Afridi b Haris Sohail	18	44	36	0	0	50.00
GC Wilson†	c Wahab Riaz b Sohail Khan	29	46	38	2	0	76.31
KJ O'Brien	c Sohaib Maqsood b Wahab Riaz	8	40	16	1	0	50.00
SR Thompson	c Umar Akmal b Rahat Ali	12	14	15	1	0	80.00
JF Mooney	c Umar Akmal b Wahab Riaz	13	27	19	1	0	68.42
GH Dockrell	run out (*Sarfraz Ahmed/Wahab Riaz)	11	13	8	0	1	137.50
AR Cusack	not out	1	3	1	0	0	100.00
Extras	(lb 2, w 10)	12					
Total	(all out; 50 overs; 216 mins)	237					(4.74 runs per over)

Match Scorecard

Ed Joyce 
Ireland

Full name Edmund Christopher Joyce

Born September 22, 1978, Dublin

Current age 36 years 175 days

Major teams England, Ireland, England Lions, Marylebone Cricket Club, Middlesex, Sussex


Nickname Joycey, Spud, Piece

Playing role Top-order batsman

Batting style Left-hand bat

Bowling style Right-arm medium

Height 5 ft 10 in



Player Profile

PlayerInfo
Id
Name
Team

The scorecard also has the link to the text commentary which is what is used to parse fielding events using keywords. In the highlighted case here Joyce is the batter involved, and using the player information table he can be uniquely identified by matching it with his full name. Each fielding event is stored and then a separate table with cumulative player fielding events is generated for each match which identifies the match and players involved uniquely.

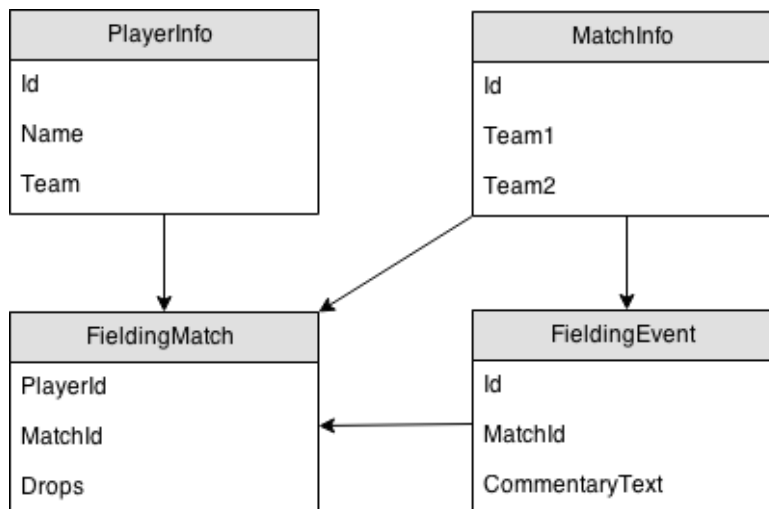
- 7.1 Rahat Ali to Joyce, no run, on a length, on off, defended off the front foot
- 7.2 Rahat Ali to Joyce, 1 run, **dropped**. Just when we thought Pakistan were up for it in the field. He has wrongfooted himself. Short and wide, cut away in the air, Shehzad moves to his left, but the ball swerves away to his right. He gets only one hand to it. A difficult chance given that movement away from him 🏏
- 7.3 Rahat Ali to Porterfield, no run, defended off the back foot
- 7.4 Rahat Ali to Porterfield, no run, fuller, on off, pushed to mid-off
- 7.5 Rahat Ali to Porterfield, **FOUR**, and once he drops ever so short, Porterfield is back and ready to pull. He has been disdainful with that shot today. Goes over midwicket this time 🏏
- 7.6 Rahat Ali to Porterfield, 1 run, goes for the big flashy square cut, gets a thick edge but it doesn't carry to third man

Match Text Commentary



FieldingEvent	FieldingMatch
Id	PlayerId
MatchId	MatchId
CommentaryText	Drops

Table Schema



Data Acquisition & Storage

In the highlighted dropped event below, the capitalized words are checked in the player table to identify the players involved. Here Buttler and Raina are both in the text, but from the initial part of the text where it says "Jadeja to Buttler" it is known that Buttler is the batter so therefore not the fielder. This way Raina is identified as the fielder who dropped the catch. This is stored in the *fieldingEvent* table, and then an aggregated version with only the fielding event counts grouped by player for the match is stored in the *fieldingMatch* table.

- 26.1 Jadeja to Buttler, **FOUR**, flighted outside off, Buttler crunches a powerful drive and beats cover to get four. It was a misfield actually from Kohli
- 26.2 Jadeja to Buttler, no run, not much room and that's driven to short cover
- 26.3 Jadeja to Buttler, no run, wider outside off and it's cut to cover point
- 26.4 Jadeja to Buttler, no run, **dropped** at slip! Buttler reverse swept from outside off, it went to Raina at slip and he put it down
- 26.5 Jadeja to Buttler, no run, outside the off stump and he tucks it to backward point
- 26.6 Jadeja to Buttler, no run, tries to cut off the back foot outside off but misses



```
select batsman, bowler, fielder, droppedCatch, commentary FROM fieldingEventODI where  
droppedCatch=1 and fielder=?,'(Suresh Raina',)
```



```
<u'Kevin Pietersen', u'Ramesh Powar', u'Suresh Raina', 1, u'dropped Pietersen comes down the track  
fairly hard chance that a costly miss')  
<u'Brian Lara', u'Ajit Agarkar', u'Suresh Raina', 1, u'oooh he almost had his man there dropped by  
ls to hold on')  
<u'Thilina Kandamby', u'Ashish Nehra', u'Suresh Raina', 1, u'almost a yorker Kandamby can't get u  
<u'Joe Root', u'Ravindra Jadeja', u'Suresh Raina', 1, u'Raina drops a dolly Root went for another  
<u'Kumar Sangakkara', u'Suresh Raina', u'Suresh Raina', 1, u'dropped by Raina Straightforward cha  
head high chance')  
<u'Mitchell Johnson', u'Ravindra Jadeja', u'Suresh Raina', 1, u'pitched up on and around off as J  
<u'Kane Williamson', u'Mohammed Shami', u'Suresh Raina', 1, u'dropped looks like Kane is in a moo  
<u'Mominul Haque', u'Umesh Yadav', u'Suresh Raina', 1, u'dropped by the India captain Instinctive  
<u'Eoin Morgan', u'Suresh Raina', u'Suresh Raina', 1, u'gives himself room and slaps a full ball  
<u'Seekkuge Prasanna', u'Umesh Yadav', u'Suresh Raina', 1, u'dropped by Raina to his left at slip  
u'Jos Buttler', u'Ravindra Jadeja', u'Suresh Raina', 1, u'dropped at slip Buttler reverse swept
```

Data Cleaning

To accurately parse the text commentary data, packages like NLTK were considered but because cricket-specific terms are not recognized and difficult to evaluate it was decided that searching for specific keywords in the text to identify events was a better way forward.

"bobbles the chance", "has made a meal of it", "sitter", "dolly", "spills", "put down", "dropping the ball", "gets both hands to it but drops it", "drops an easy catch", "fails to take the catch", "dropped", "shelled", "grassed"

Dropped Catch Key Words

"dolly on the toes", "dropped right", "dropped with", "dropped just", "dropped it short", "dropped short", "dropped well in front", "drops the wrist", "dropped from", "earlier he was dropped", "dropped his", "dropped a touch short", "dropped catches", "dropped behind", "dropped at his feet", "dropped in", "dropped a bit", "dropped into", "dropped softly", "dropped his bat", "dropped catch and", "dropped it into", "dropped to", "dropped him earlier", "dropped far too short", "dropped over", "tough chance", "hard chance", "hard to call it dropped", "hard to call that dropped", "like a football goalkeeper", "desperate effort", "difficult chance", "superb attempt", "good effort", "screaming past", "would have been a very good", "terrific effort", "harsh to blame", "great attempt", "what an effort", "would have been a terrific catch", "would have been a wundercatch", "tough one", "fabulous attempt", "tremendous effort", "difficult one", "would have been a stunner", "would have been a superb", "would have been a mind-blowing", "valiant effort", "harsh to call it", "would have been a classic catch", "would have been a cracker"

Dropped Catch Invalidating Key Words

The dropped catch fielding event was difficult to accurately evaluate because the keyword "dropped" is used in different ways in cricket. Sometimes it constitutes a dropped catch, but sometimes it might be the batter just "dropping" the ball to the ground or the pitcher "dropping" the ball short (which is a term used to signify that the pitcher has thrown the ball midway through the bases). To distinguish between different types of "dropped" keywords, an additional list of keywords that invalidated the keyword match was utilized - as an example if the text contained "dropped" but the next word was "dropped short", that case is ignored as a fielding non-event. For dropped catches there is the additional complication in that some catches are not expected to be taken because of high difficulty. Specific keywords such as "great effort" or "difficult chance" are used to ignore dropped catches for those cases.

"direct hit", "accurate with the throw", "throw has beaten him", "hits the stumps direct"

Direct Hit Key Words

This complication in keywords is not as much an issue for direct hits or great catches - this is reflected in the keywords list length being smaller for those events.

Implementation

Technologies Used

SQL vs NoSQL:

SQLite was chosen over potential NoSQL solutions like MongoDB or MarkLogic because there were no compelling reasons not to go with a relational database structure. Data is gathered from completed matches in the past and since this is an evaluation of a player's fielding ability, real-time data is not necessary. Fielding event types do not change so it is possible to start with exhaustive table schema that does not require modification in the future. All the scraped data size amounts to around 150MB which is sizeable but not massive - it also grows steadily since the number of matches played in a year is consistent and does not change significantly.

Local vs EC2:

In terms of running the parsing scripts – running locally made most sense instead of using EC2 for more servers since the whole scraping running time to parse scorecard details and text commentary is around 10 hours. After a complete run is completed only newer matches would need be scraped – this would not take more than 5 minutes at a time when run locally.

PHP vs Other:

PHP, HTML and CSS were used to display the results of the project in accessible web form. PHP has built-in support to connect to SQLite databases, and HTML/CSS is adequate web design tools for the purposes of this project.

Description of Scripts

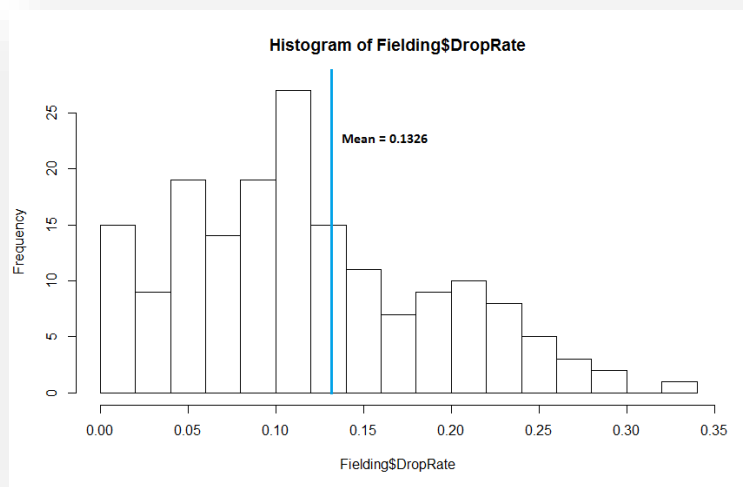
- *createTables.py*: Creates required tables in SQLite which are used in all scripts of project
- *dumpMatchInfo.py*: Scrapes match results lists for match details on teams involved and scorecard links
- *scrapeScorecard.py*: Scrapes match scorecard for batters and bowlers involved in relevant match
- *scrapeFielding.py*: Scrapes text commentary feed of match and parses fielding events and players involved – matching them with existing player table
- *dumpFieldingCareer.py*: Aggregates fielding event data by match over player careers for meaningful statistics
- *cricStats.db*: SQLite database of generated data
- *cricStats.php*: PHP webpage for appropriate display of results
- *style.css*: Styling for webpage

Results

Metrics

Once fielding events are parsed for all matches with the fielders identified, it is possible to generate meaningful statistics to enable comparison between players. One revealing metric is drop rate - the percentage chance that a fielder would drop a catch that he was supposed to have taken.

$$\text{Drop \%} = \frac{\text{Dropped Catches} - \text{Difficult Chances}}{(\text{Catches} + \text{Dropped Catches} - \text{Difficult Chances})}$$



The histogram above shows that the mean drop rate is 13.26% with the graph showing positive skew – indicating that a majority of players perform around or better than the mean drop rate, with a minority of poor fielders with higher drop rates. A fielder with a drop rate less than 10% can be considered to be above average just based on catches.

Rank	Player	Country	Span	Mat	Catches	Drops	Drop %	Great Catches	Direct Hits	Saves	Runs
1	Ricky Ponting		2005–2012	68	62	8	11.43	5	10	4	–8
2	Michael Hussey		2005–2012	68	75	5	6.25	8	1	3	0
3	Paul Collingwood		2002–2011	62	68	8	10.53	10	1	5	–1
4	AB de Villiers		2006–2015	105	144	8	5.26	3	5	6	–1
5	Denesh Ramdin		2006–2015	73	121	5	3.97	2	2	2	–2
6	Kumar Sangakkara		2005–2015	182	278	18	6.08	10	2	2	–2
7	Adam Gilchrist		2002–2008	56	100	3	2.91	2	0	3	3
8	Dwayne Bravo		2006–2014	66	67	7	9.46	7	2	5	2
9	Martin Guptill		2009–2015	50	49	3	5.77	2	2	10	10
10	Steven Smith		2010–2015	28	31	2	6.06	1	5	2	–1

The table above displays the top fielders rated based on different fielding attributes. Almost all of their drop rates are below 10%, with those that have higher rates like Ricky Ponting and Paul Collingwood having other positive attributes such as high direct hit rates or high great catch rates. Ricky Ponting, Paul Collingwood and AB de Villiers are all considered great fielders – this provides validation to the data and suggests that it is on the right track.

Challenges

The most obvious challenge with this method of parsing fielding events is that it is impossible to guarantee 100% accuracy. Sometimes there are more than 2 players involved in the text and it would be impossible to identify who the correct fielder is without manual observation. Misattribution becomes an issue that is hard to avoid. Even so, based on manual comparisons of a sample of matches, the parsing method has a 90%+ accuracy.

Another issue is that this setup is relatively inflexible and relies on *cricinfo*. If the site changes its format, the parsing infrastructure would also require modification. This also points to long-term viability issues.

Scalability

The current setup focuses on One-day International (ODI) cricket which is one of three forms of the game. The infrastructure used here can be mimicked for Test and Twenty20 cricket with minor modifications – allowing for a three-fold increase in generated data rapidly.

Conclusion

Analysis of Result

The project was successful in achieving the stated goal of generating cricket fielding statistics. Although perfect accuracy proved impossible to guarantee, the results were still useful to a great degree and could be used to evaluate player ability as it is – especially considering there are no other current measures in the sport.

Improvements or Alternatives

There are no obvious possible improvements that would solve misattribution issues without manual intervention – so the only plausible better alternative would be to implement a video analysis infrastructure in the game.

Future Iterations

Even though this project successfully generated fielding statistics such as drop rate, it is still difficult to quantify how much value fielding brings to the game. If a player has a 50% drop rate while averaging 40 runs a game with the bat is he better for the team than someone with a 10% drop rate but averages just 30?

1st innings			
Over	Score	Runs	Run Rate
1	8/0	8	8.00
2	9/0	1	4.50
3	17/0	8	5.66
4	30/0	13	7.50
5	36/1	6	7.20
6	48/2	12	8.00
7	54/2	6	7.71
8	63/2	9	7.87
9	68/2	5	7.55
10	73/3	5	7.30
11	82/3	9	7.45
12	87/4	5	7.25
13	89/5	2	6.84
14	93/5	4	6.64
15	100/5	7	6.66
16	109/5	9	6.81
17	119/5	10	7.00
18	122/7	3	6.77
19	127/8	5	6.68
20	130/9	3	6.50

To value batting, bowling and fielding impact more accurately, a future iteration could use a "Win Percentage Added" (WPA) measure.

For example - in the table to the left a particular team's batting performance is displayed. An over is 6 pitches - and this match has 20 overs. The score column shows that at the 11th over the team was 82/3 - which means they had 82 runs with 3 outs. After the 12th over, the score is 87/4, which means the batting team got 5 runs, but also lost a batter.

It is possible to parse these over-score tables for all historical matches and look for similar scores at the 11th and 12th overs historically. This allows for win odds for the batting team to be generated at each stage of the match.

A possible scenario is that the batting team has 55% win odds after the 11th over, but after the loss of a wicket it goes down to 50% after the 12th. Ignoring the runs scored, if that out was caused by a direct hit from a fielder, he is attributed 5% in win percentage added. This method allows for more accurate measurement of player impact – in fielding as well as batting and bowling.