

Validating Web-Based MIDI Content for Digital Libraries

Daniel Shanahan, Louisiana State University
Joshua Albrecht, University of Mary Hardin Baylor

October 3, 2014 Baton Rouge

Validating Web-Based MIDI Content for Digital Libraries

Daniel Shanahan, Louisiana State University
Joshua Albrecht, University of Mary Hardin Baylor

October 3, 2014 Baton Rouge

Validating Web-Based MIDI Content for Digital Libraries

Daniel Shanahan, Louisiana State University
Joshua Albrecht, University of Mary Hardin Baylor

October 3, 2014 Baton Rouge

Background

- Bronson (1949)
 - IBM Card Reader
- *Josquin Project* (Hall, 1975)
 - Princeton project, established a network of copied manuscript sources.
- Lomax's *Cantometrics* (1968)
 - Computational and statistical approaches to “world music.”
- *Darms Project* (Bauer-Mengelberg, discussed in Erickson, 1976)
 - Comprehensive encoding language for representing notating music.

Background

- Bronson (1949)
 - IBM Card Reader
- *Josquin Project* (Hall, 1975)
 - Princeton project, established a network of copied manuscript sources.
- Lomax's *Cantometrics* (1968)
 - Computational and statistical approaches to “world music.”
- *Darms Project* (Bauer-Mengelberg, discussed in Erickson, 1976)
 - Comprehensive encoding language for representing notating music.

Existing Datasets

- Kernscores (CCARH)
- RISM
- Barlow and Morgenstern (Huron, 1996)
- Essen Folksong Collection (Schaffrath, 1995)
- Billboard Corpus of Rock Songs (Burgoyne, 2012)
- Jazz Standards Dataset (Shanahan and Broze, 2012)
- Densmore Collection of Native American Folksongs (Shanahan and Shanahan, 2014)

Existing Datasets

- Kernscores (CCARH)
- RISM
- Barlow and Morgenstern (Huron, 1996)
- Essen Folksong Collection (Schaffrath, 1995)
- Billboard Corpus of Rock Songs (Burgoyne, 2012)
- Jazz Standards Dataset (Shanahan and Broze, 2012)
- Densmore Collection of Native American Folksongs (Shanahan and Shanahan, 2014)

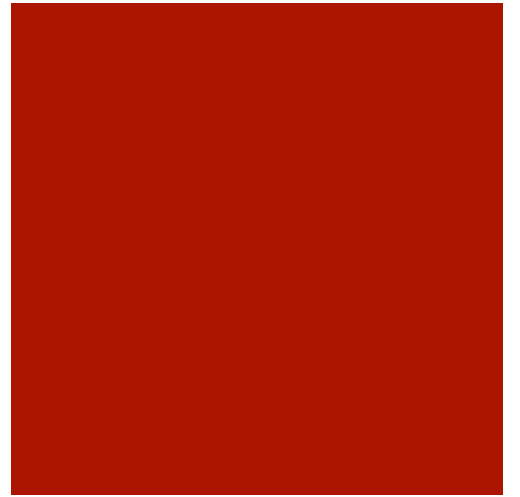
Existing Datasets

- Kernscores (CCARH)
- RISM
- Barlow and Morgenstern (Huron, 1996)
- Essen Folksong Collection (Schaffrath, 1995)
- Billboard Corpus of Rock Songs (Burgoyne, 2012)
- Jazz Standards Dataset (Shanahan and Broze, 2012)
- Densmore Collection of Native American Folksongs (Shanahan and Shanahan, 2014)

Existing Datasets

- Kernscores (CCARH)
- RISM
- Barlow and Morgenstern (Huron, 1996)
- Essen Folksong Collection (Schaffrath, 1995)
- Billboard Corpus of Rock Songs (Burgoyne, 2012)
- Jazz Standards Dataset (Shanahan and Broze, 2012)
- Densmore Collection of Native American Folksongs (Shanahan and Shanahan, 2014)

The Corpus Study Dilemma...



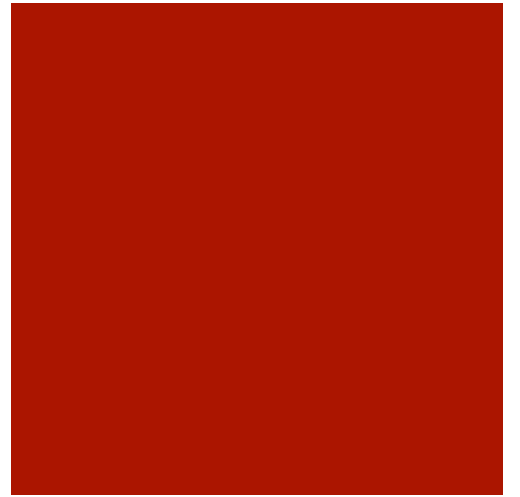
- The tools for analysis are there, but we've been using the same datasets for decades.
 - (encoding is time consuming)
- Repeated use of same datasets
 - (Bach Chorales, Essen Folksongs, etc.)

The Corpus Study Dilemma...



- Datasets currently available are quite canonical (Bach, Mozart, Haydn, and Beethoven make up a large portion of manually encoded corpora).
- A lack of compositions from outside the Common Practice Era (although this is changing!).
 - Even with these composers, the repertoire is quite limited:
 - There are relatively few large orchestral works or operas.
 - Encoding piano works or small-scale works is obviously preferred to encoding symphonies.

The MIDI Solution



- One possible solution might be to re-examine MIDI data.
- With MIDI, we have exponentially more symbolic data than we do from the various encoding initiatives combined.

The Nature of MIDI Data and XML Conversion

28.5,	NOTE,	0.09375,	69,	85,	0,	0;
28.5,	NOTE,	0.09375,	42,	77,	0,	0;
28.5,	NOTE,	0.09375,	54,	77,	0,	0;
28.5,	NOTE,	0.09375,	66,	85,	0,	0;
28.5,	NOTE,	0.09375,	73,	85,	0,	0;
29,	NOTE,	0.635417,	38,	77,	0,	0;
29,	NOTE,	0.635417,	78,	85,	0,	0;
29,	NOTE,	0.635417,	71,	85,	0,	0;
29,	NOTE,	0.635417,	69,	85,	0,	0;
29,	NOTE,	0.635417,	66,	85,	0,	0;
29,	NOTE,	0.635417,	50,	77,	0,	0;
30,	CONT,	64,	127,	-1000,	0,	0;
30,	CONT,	64,	127,	-1000,	0,	0;
30,	NOTE,	0.125,	40,	77,	0,	0;
30.125,	NOTE,	0.125,	44,	77,	0,	0;
30.25,	NOTE,	0.125,	47,	77,	0,	0;
30.375,	NOTE,	1.26042,	52,	77,	0,	0;
30.5,	NOTE,	0.125,	59,	85,	0,	0;
30.625,	NOTE,	0.125,	62,	85,	0,	0;
30.75,	NOTE,	0.125,	64,	85,	0,	0;
30.875,	NOTE,	0.760417,	68,	85,	0,	0;
32,	TEMPO,	81.0001,	-1000,	-1000,	-1000,	-1000;
32,	CONT,	64,	0,	-1000,	0,	0;
32,	CONT,	64,	0,	-1000,	0,	0;
32,	CONT,	64,	127,	-1000,	0,	0;
32,	NOTE,	0.125,	33,	-	-	-
32.125,	NOTE,	0.125,	37,	-	-	-
32.25,	NOTE,	0.125,	40,	-	-	-
32.375,	NOTE,	0.9375,	45,	-	-	-
32.5,	CONT,	64,	127,	-	-	-
32.5,	NOTE,	0.125,	57,	-	-	-
32.625,	NOTE,	0.125,	61,	-	-	-
32.75,	NOTE,	0.125,	64,	-	-	-
32.875,	NOTE,	0.4375,	69,	-	-	-
33.5,	NOTE,	0.0729167,	-	-	-	-
33.625,	NOTE,	0.0729167,	-	-	-	-
33.75,	NOTE,	0.166667,	-	-	-	-
33.9167,	NOTE,	0.166667,	-	-	-	-
34.0833,	NOTE,	0.166667,	-	-	-	-
34.25,	NOTE,	0.166667,	-	-	-	-
34.4167,	NOTE,	0.166667,	-	-	-	-
34.5833,	CONT,	64,	-	-	-	-
34.5833,	NOTE,	0.166667,	-	-	-	-
34.75,	NOTE,	0.5625,	33,	-	-	-
35,	TEMPO,	40,	-1000,	-	-	-
35.5,	TEMPO,	81.0001,	-	-	-	-

2 Flauti

2 Oboi

Clarinetti in B

2 Fagotti

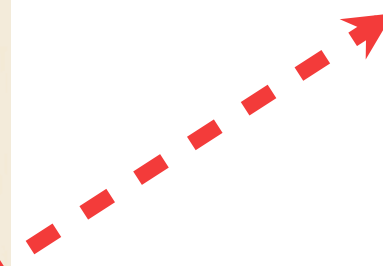
Corno I in B

Corno II in B

Violine I

Violine II

```
**kern **kern
*staff2 *staff1
=1- =1-
*clefF4 *clefG2
*k[b-e-a-d-g-c-] *k[b-e-a-d-g-c-]
*M3/4 *M3/4
2.r 4r
. 4r
. 4B-/
=2 =2
8EE-/L 4.g-/
8BB-/J .
4E-\ 8G-/
4r 8B-/L 8a-/
. 8e-/J 8g-/
=3 =3
2.DD-/ 8G-/L 8B-/ 8f/ 8g-/
. 8e-/J
. 4b-\
. 4g-/
=4 =4
8CC-/L 4ee-\
8C-/J 8G-/L
4r 8e-/J
. 8g-/L 8ff/
. 8A-/J 8a-/ 8ee-/
=5 =5
2.GGG-/ 2.GG-/ 8B-/L 8g-/ 8b-/ 8dd-/
. 8cc-/J
. 4dd-\
. 4g-/
=6 =6
```



2 Flauti

2 Oboi

Clarineti in B

2 Fagotti

Corno I in B

Corno II in B

Violine I

Violine II

Viola

Violoncelle

Contrabbasse



4



2 Flauti

2 Oboi

Clarineti in B

2 Fagotti

Corno I in B

Corno II in B

Violine I

Violine II

Viola

Violoncelle

Contrabbasse

A musical score for the first system of an orchestra. The score includes parts for 2 Flauti, 2 Oboi, Clarineti in B, 2 Fagotti, Corno I in B, Corno II in B, Violine I, Violine II, Viola, Violoncelle, and Contrabbasse. The key signature is B-flat major (two flats) and the time signature is 4/4. The first system shows the initial measures of the piece, with most instruments playing sustained notes or rests.

4

A musical score for the second system of an orchestra. The score continues from the first system, showing measures 4 through 6. The instrumentation remains the same. The second system shows more active musical development, with various instruments playing moving lines and chords.

2 Flauti

2 Oboi

2 Clarinetti in B

2 Fagotti

Corno I in B

Corno II in B

Violine I

Violine II

Viola

Violoncelle

Contrabbasse

Allegretto scherzando (♩ = 88)

2 Flauti

2 Oboi

2 Clarinetti in B

2 Fagotti

2 Corni in B

Violino I

Violino II

Viola

Violoncello e Contrabbasso

pp sempre staccato

pp

pizz.

pp

pizz.

p

pp

Chopin: Op. 24 No. 3

MusicXML Part



The first system of musical notation for Chopin's Op. 24 No. 3, measures 1-2. The treble clef staff contains a melodic line with eighth and sixteenth notes, including a triplet of eighth notes. The bass clef staff contains a supporting line with eighth notes and a triplet of eighth notes. A bracket labeled '6' is placed under the first triplet in the bass staff. A green square button with a left-pointing arrow is located in the top right corner.

The second system of musical notation for Chopin's Op. 24 No. 3, measures 3-4. The treble clef staff begins with a measure rest labeled '2' and contains a melodic line with eighth and sixteenth notes, including a triplet of eighth notes. The bass clef staff contains a supporting line with eighth notes and a triplet of eighth notes. A bracket labeled '7' is placed under the first triplet in the bass staff. A green square button with a left-pointing arrow is located in the top right corner.

The third system of musical notation for Chopin's Op. 24 No. 3, measures 5-6. The treble clef staff begins with a measure rest labeled '3' and contains a melodic line with eighth and sixteenth notes, including a triplet of eighth notes. The bass clef staff contains a supporting line with eighth notes and a triplet of eighth notes. A bracket labeled '5' is placed under the first triplet in the bass staff. A green square button with a left-pointing arrow is located in the top right corner.

Chopin: Op. 24 No. 3

MusicXML Part

First system of musical notation for Chopin's Op. 24 No. 3. It consists of a treble and bass staff in 4/4 time. The treble staff has a melodic line with eighth and sixteenth notes. The bass staff has a supporting line with eighth notes and rests. There are two green square icons with a left-pointing arrow in the top right corner of the system.

Second system of musical notation for Chopin's Op. 24 No. 3. It consists of a treble and bass staff in 4/4 time. The treble staff has a melodic line with eighth and sixteenth notes. The bass staff has a supporting line with eighth notes and rests. There are two green square icons with a left-pointing arrow in the top right corner of the system.

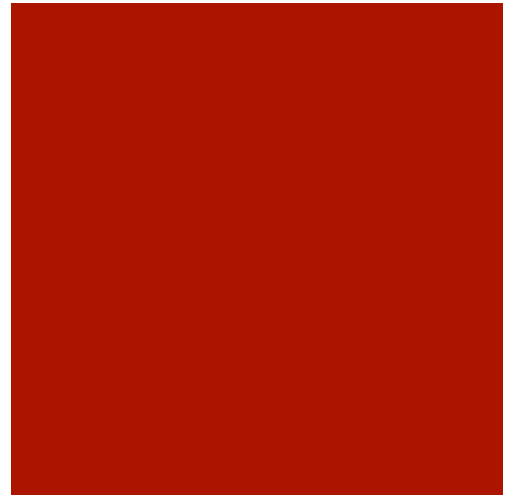
Third system of musical notation for Chopin's Op. 24 No. 3. It consists of a treble and bass staff in 4/4 time. The treble staff has a melodic line with eighth and sixteenth notes. The bass staff has a supporting line with eighth notes and rests. There are two green square icons with a left-pointing arrow in the top right corner of the system.

Fourth system of musical notation for Chopin's Op. 24 No. 3. It consists of a treble and bass staff in 3/4 time. The treble staff has a melodic line with eighth and sixteenth notes. The bass staff has a supporting line with eighth notes and rests. There are two green square icons with a left-pointing arrow in the top right corner of the system.

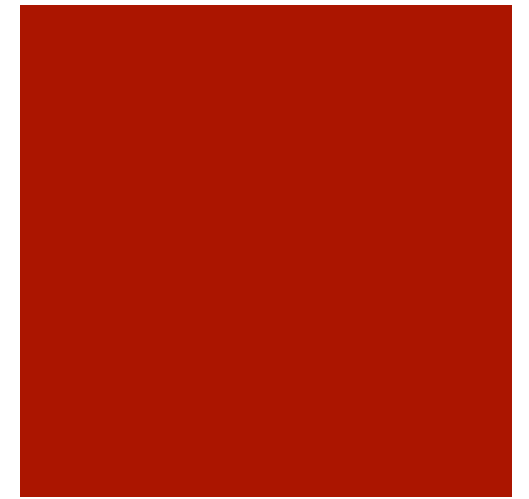
Fifth system of musical notation for Chopin's Op. 24 No. 3. It consists of a treble and bass staff in 3/4 time. The treble staff has a melodic line with eighth and sixteenth notes. The bass staff has a supporting line with eighth notes and rests. There are two green square icons with a left-pointing arrow in the top right corner of the system.

Motivation

- Two Arguments:
 - MIDI data might be somewhat flawed, but the amount of data can allow us to see a signal through the noise.
 - MIDI data are inherently messy, and it would be unwise to use the medium in a corpus study.

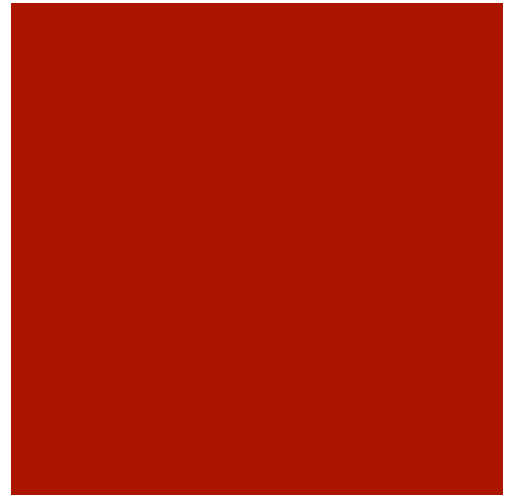


What can we get from MIDI data?



- Just how accurate are these online MIDI datasets?
- Which questions can we be asking, that might provide reasonably reliable results?

Validation



- We decided to validate a subset of Classical Archives MIDI dataset.
 - Now being used as the “Yale/Classical Archives MIDI Dataset”
- We analyzed the following types of errors:
 - Pitch, rhythm (onset and duration), and orchestration
 - Unfortunately, form markers (repeats, section markers, etc) are lost in translation, and everything is seen as being through-composed.

Classical Archives

More than 12,000 MIDI files from a number of sources, most of which were submissions by users.

Most popular tracks now ▶ Christmas Oratorio (6 Cantatas, from Christmas to Epiphany), BWV248 • Bruch - Violin Concerto No.1 in G-, Op.26 • Bartók - Hungarian Peas

ClassicalArchives™
THE ULTIMATE CLASSICAL MUSIC DESTINATION

0 Items — **VIEW CART** ▶

Search for Music

ADVANCED SEARCH ▶

HOME **PLAY MUSIC** **DOWNLOAD** **MIDI** **YOUR ACCOUNT**

By Composer ▶ **By Contributor** ▶

Read what our subscribers say!

Use Facebook login

OR USE MEMBER LOGIN ▶

Send a Personalized **GIFT CERTIFICATE**

FREE TRIAL

Log-in now or start your **FREE 14-day trial membership** for UNLIMITED PLAY.

Home ▶ Composers (MIDI) ▶ G ▶ Charles Gounod

Composer (MIDI)

Charles Gounod (1818-1893); FRA

ABOUT/BIO **WORKS**

 Charles Gounod is best known for his operas Faust and Romeo et Juliette and for his Ave Maria (1859). Except... **More** ▶

These synthesizer sequences are free to play or download by all logged-in members.

MIDI: 41 midis  Click on a category to view the list of works

- [Stage Works](#)
- [Operas](#)
- [Ballets](#)
- [Vocal Works](#)
- [Songs \(includes Bach-Gounod, Ave Maria\)](#)
- [Choral Works](#)
- [Orchestral Works](#)
- [Symphonies](#)
- [Other Orchestral Works](#)
- [Chamber Works](#)

• Stage Works 27 midis **TOP** ▲

- Operas 19 midis
- ▶ [Faust \(opéra\)](#) 19 midis
- Ballets 8 midis **TOP** ▲
- ▶ [Faust \(ballet, based on the opera\)](#) 8 midis
- Vocal Works 9 midis **TOP** ▲

NEW!!!
CLASSICAL ARCHIVES
ANYWHERE, ANYTIME!


Available on the **App Store**

GET IT ON **Google play**  Available at **amazon appstore** for Android

Subscribe and enjoy these free apps now!
Unlimited access on your phone or tablet!

ClassicalArchives

LISTEN INSTANTLY
IN FULL FROM THE
LARGEST AND BEST ORGANIZED
CLASSICAL MUSIC COLLECTION



Full works and albums
with just 1-click!

START MY FREE 14-DAY TRIAL!

Cancel anytime - Secure; no cc data retained.

Classical Archives

More than 12,000 MIDI files from a number of sources, many of which were submissions by users.

Most popular tracks now ▶ Christmas Oratorio (6 Cantatas, from Christmas to Epiphany), BWV248 • Bruch - Violin Concerto No.1 in G-, Op.26 • Bartók - Hungarian Peas

ClassicalArchives™
THE ULTIMATE CLASSICAL MUSIC DESTINATION

0 Items — **VIEW CART** ▶

Search for Music

ADVANCED SEARCH ▶

HOME **PLAY MUSIC** **DOWNLOAD** **MIDI** **YOUR ACCOUNT**

Read what our subscribers say!

Use Facebook login

OR USE MEMBER LOGIN ▶

By Composer ▶ **By Contributor** ▶

Send a Personalized GIFT CERTIFICATE

FREE TRIAL

Log-in now or start your **FREE 14-day trial membership** for UNLIMITED PLAY.

Home ▶ Composers (MIDI) ▶ G ▶ Charles Gounod

Composer (MIDI)

Charles Gounod (1818-1893); FRA

ABOUT/BIO **WORKS**

 Charles Gounod is best known for his operas Faust and Romeo et Juliette and for his Ave Maria (1859). Except... **More** ▶

These synthesizer sequences are free to play or download by all logged-in members.

MIDI: 41 midis  Click on a category to view the list of works

- [Stage Works](#)
- [Operas](#)
- [Ballets](#)
- [Vocal Works](#)
- [Songs \(includes Bach-Gounod, Ave Maria\)](#)
- [Choral Works](#)
- [Orchestral Works](#)
- [Symphonies](#)
- [Other Orchestral Works](#)
- [Chamber Works](#)

• Stage Works 27 midis **TOP** ▲

- Operas 19 midis
- ▶ [Faust \(opéra\)](#) 19 midis
- Ballets 8 midis **TOP** ▲
- ▶ [Faust \(ballet, based on the opera\)](#) 8 midis
- Vocal Works 9 midis **TOP** ▲

NEW!!! CLASSICAL ARCHIVES ANYWHERE, ANYTIME!

Available on the App Store

GET IT ON Google play

Available on amazon appstore for Android

Subscribe and enjoy these free apps now!
Unlimited access on your phone or tablet!

ClassicalArchives

LISTEN INSTANTLY IN FULL FROM THE LARGEST AND BEST ORGANIZED CLASSICAL MUSIC COLLECTION



Full works and albums with just 1-click!

START MY FREE 14-DAY TRIAL!

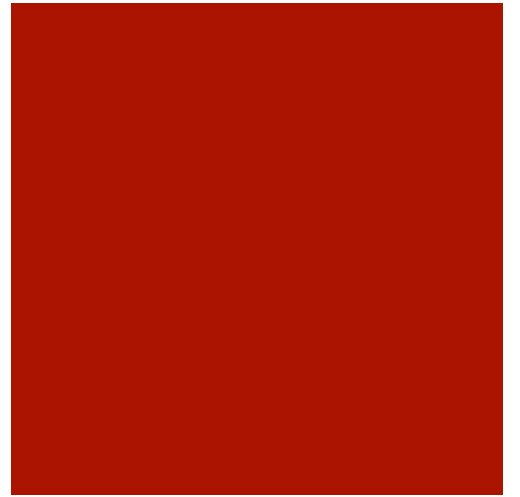
Cancel anytime - Secure: no cc data retained.

Sampling

- 100 pieces were randomly assigned to each author.
- Each author examined two consecutive onsets, which were chosen from a randomly selected beat on a randomly selected measure of the piece.
- The converted XML score was compared against a published score on the International Music Score Library Project (www.IMSLP.org).
- 7 pieces of the 200 were discarded (3 incomplete scores, 4 were mistakenly validated against IMSLP scores derived from CCARH encodings).
- Each of the remaining selections was validated through the use of 20 criteria, falling into four groups.



Pitch Errors

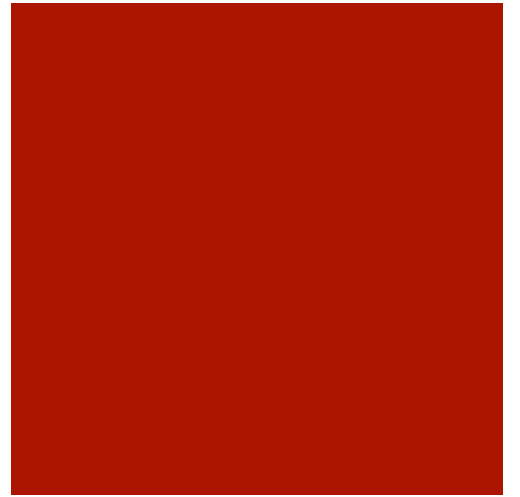


Incorrect pitch class (Onset 1)

Incorrect pitch class (Onset 2)

- If there are pitch tokens in the XML file that do not occur in the published score, mark it as incorrect. Octave errors count as errors , but enharmonic respellings do not count.

Pitch Errors

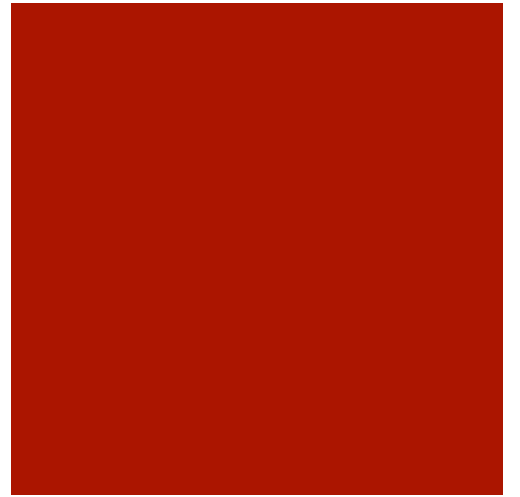


Incorrect pitch class (Onset 1)

Incorrect pitch class (Onset 2)

- If there are pitch tokens in the XML file that do not occur in the published score, mark it as incorrect. Octave errors count as errors , but enharmonic respellings do not count.

Pitch Errors

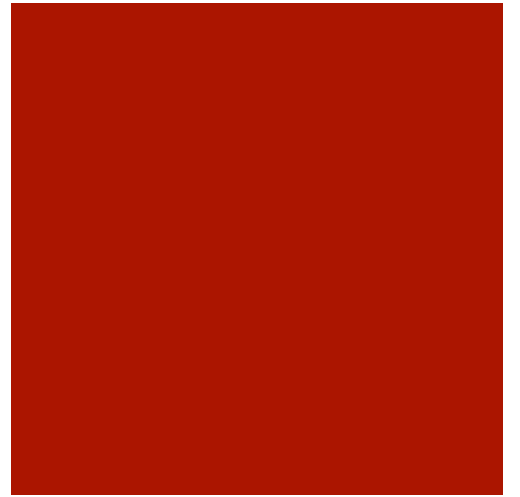


Pitch class missing (Onset 1)

Pitch class missing (Onset 2)

- If there are pitch tokens in the published score that do not occur in the XML file, mark it as incorrect.

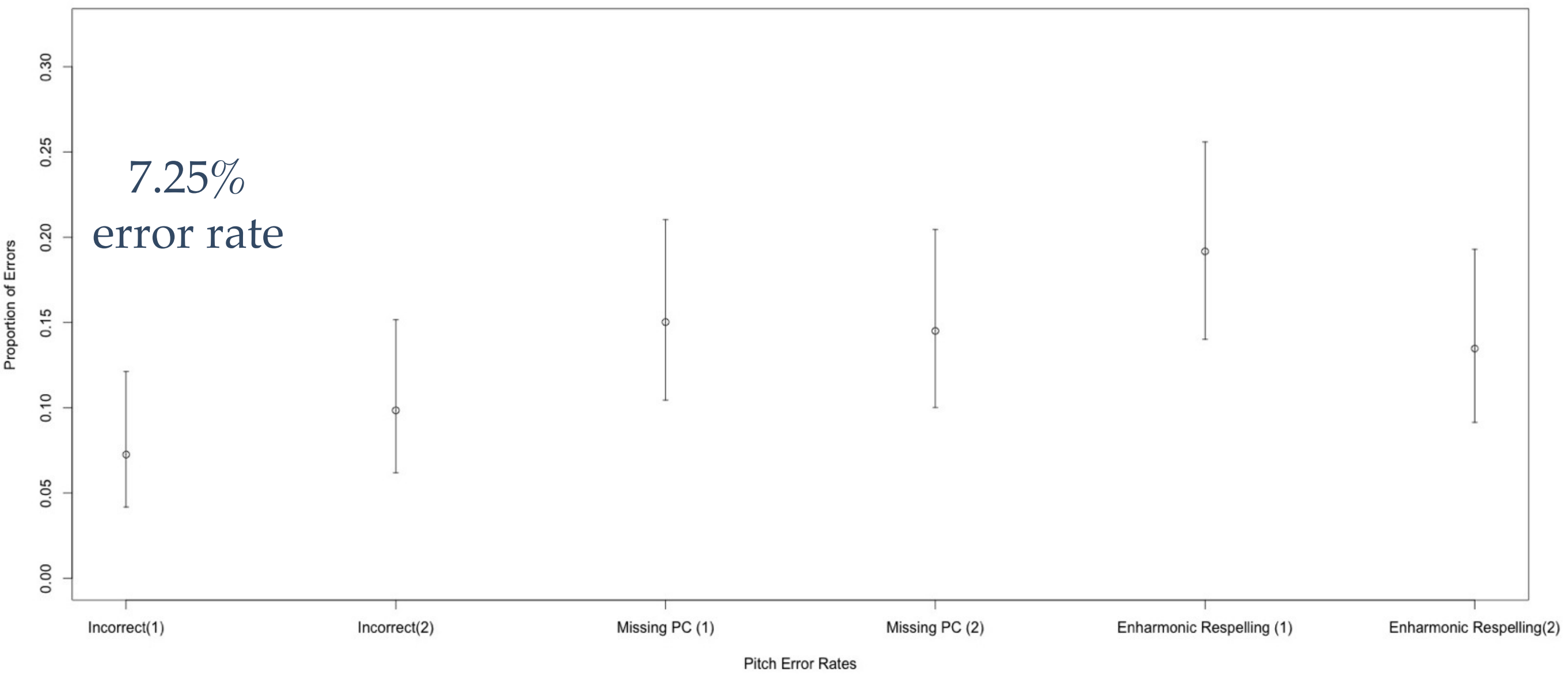
Pitch Errors

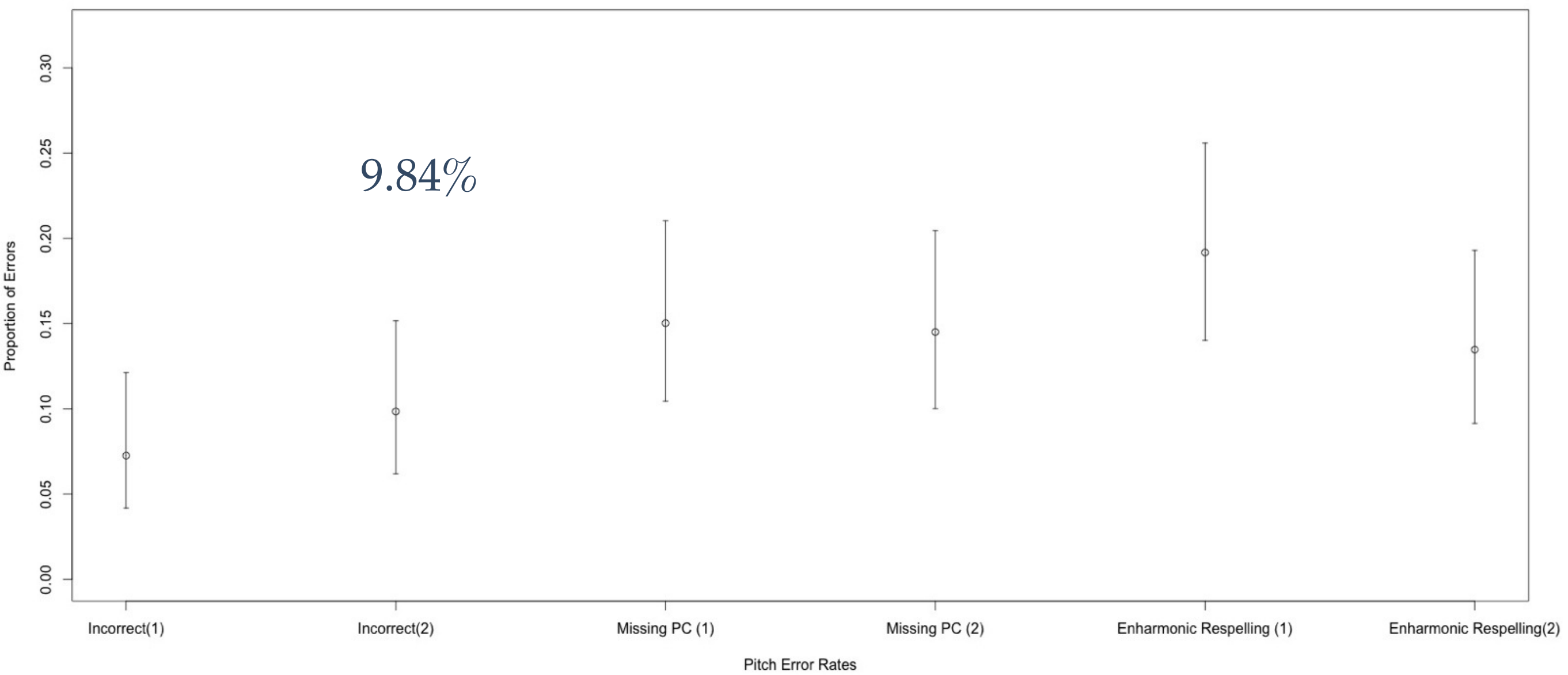


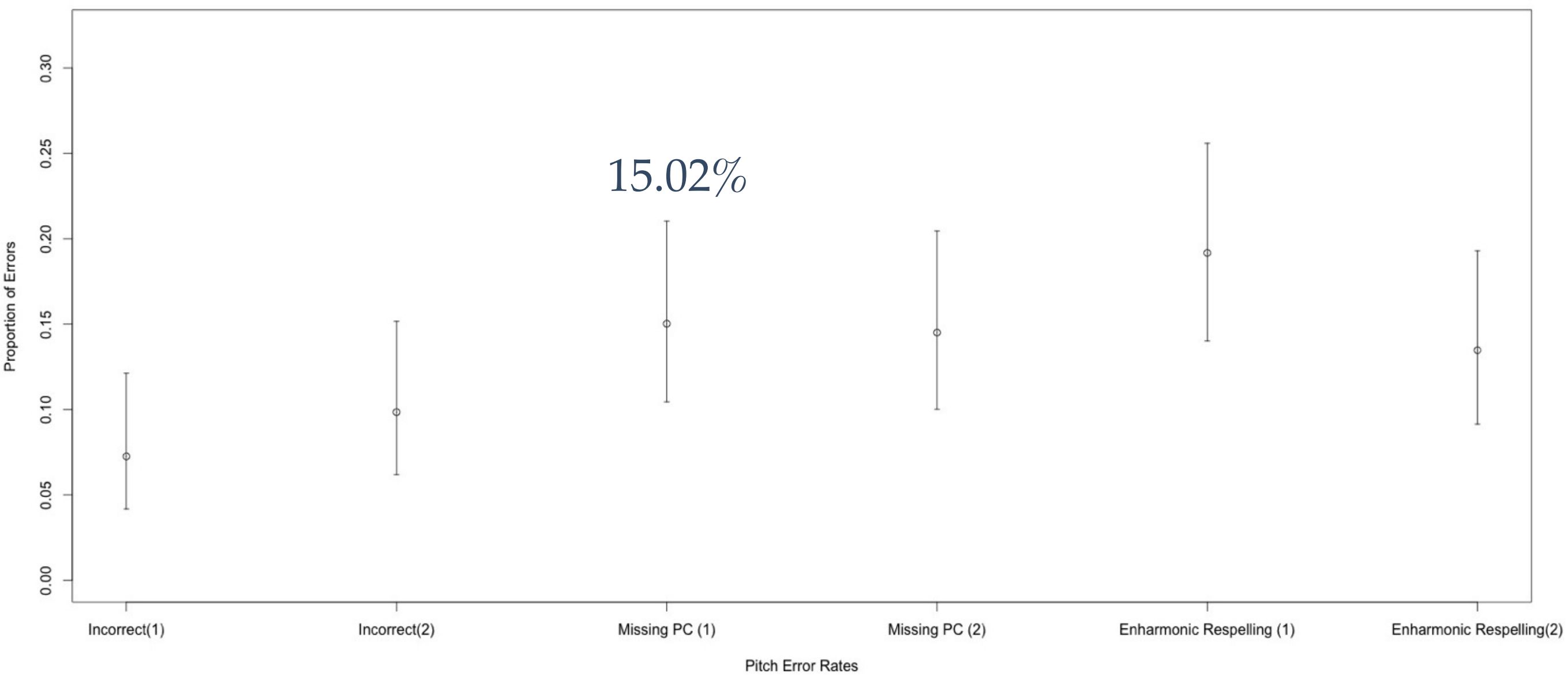
Enharmonic respelling (Onset 1)

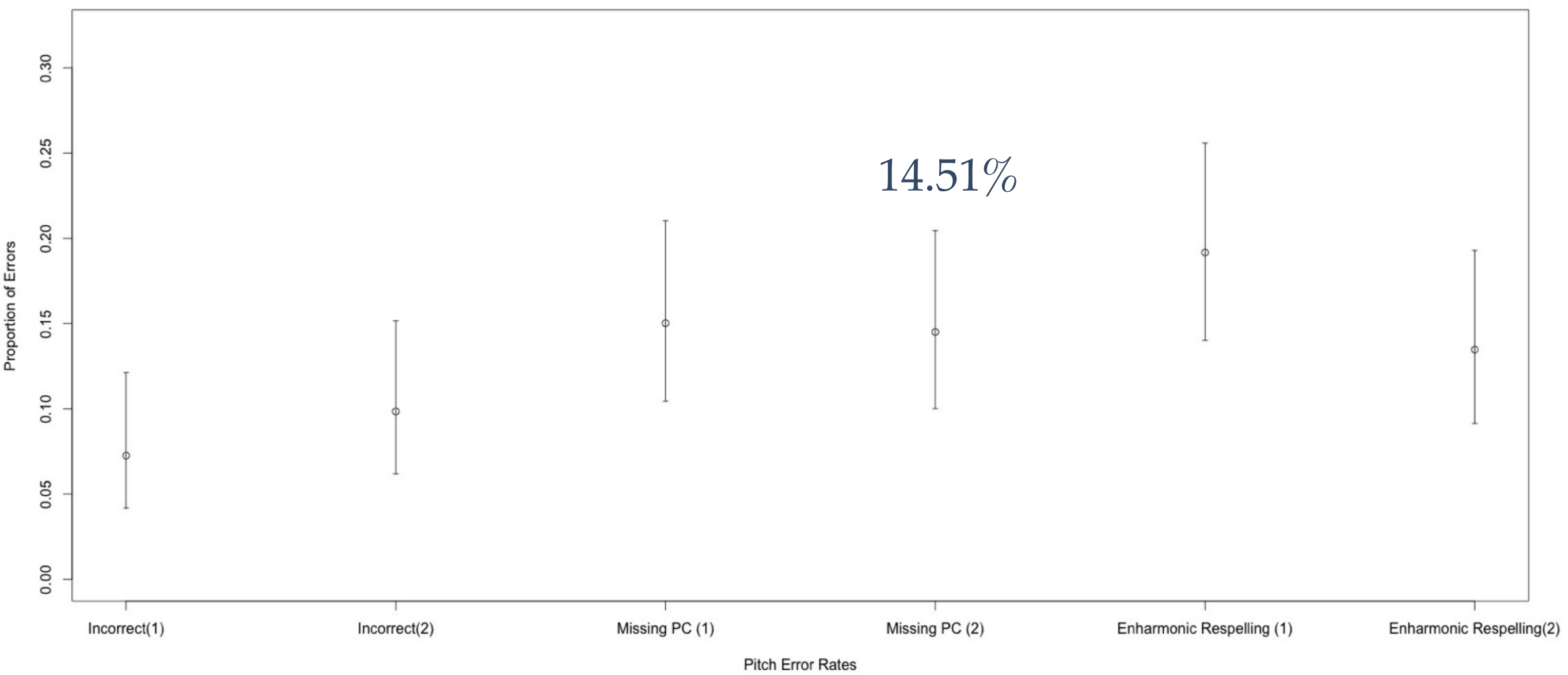
Enharmonic respelling (Onset 2)

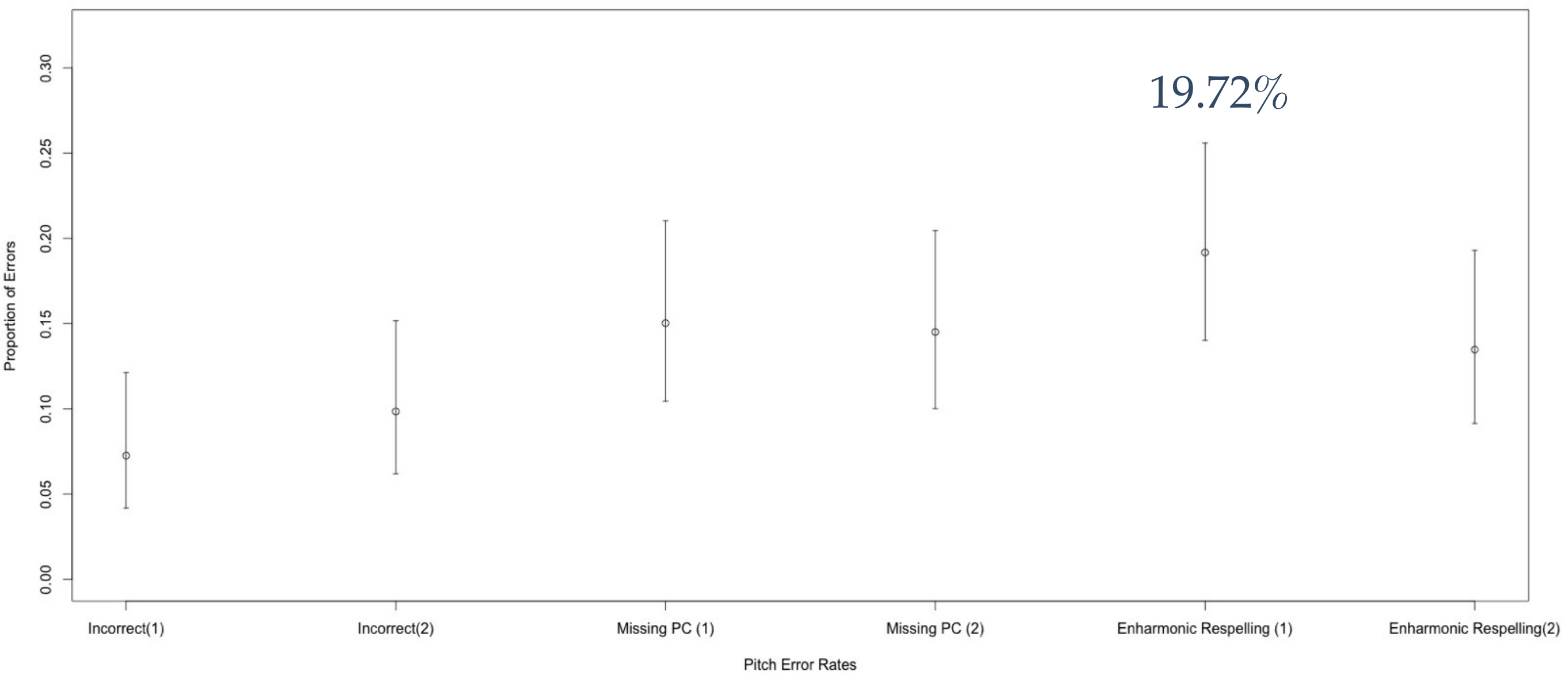
- If any of the pitches are enharmonically respelled, count as one error.

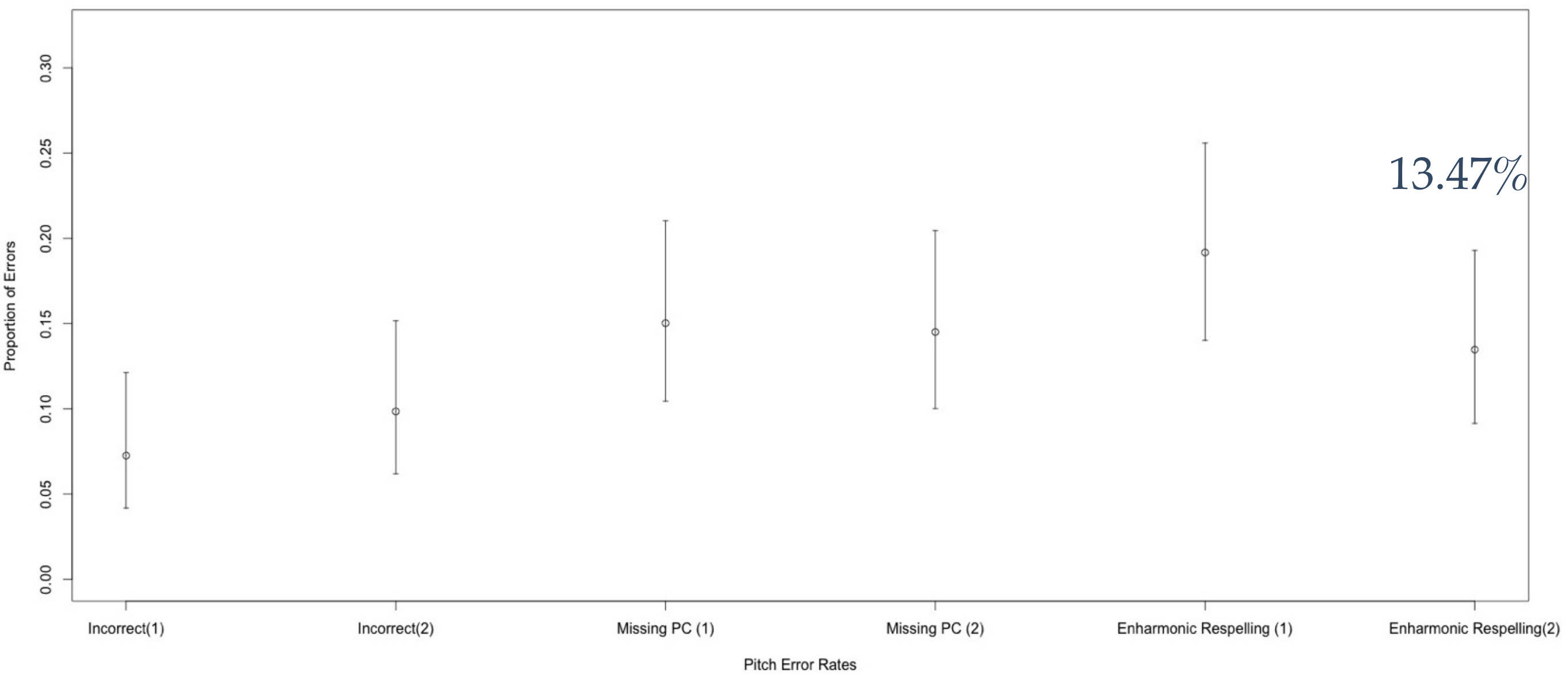




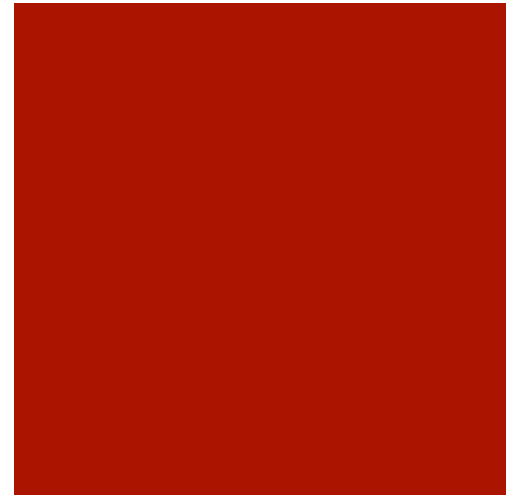




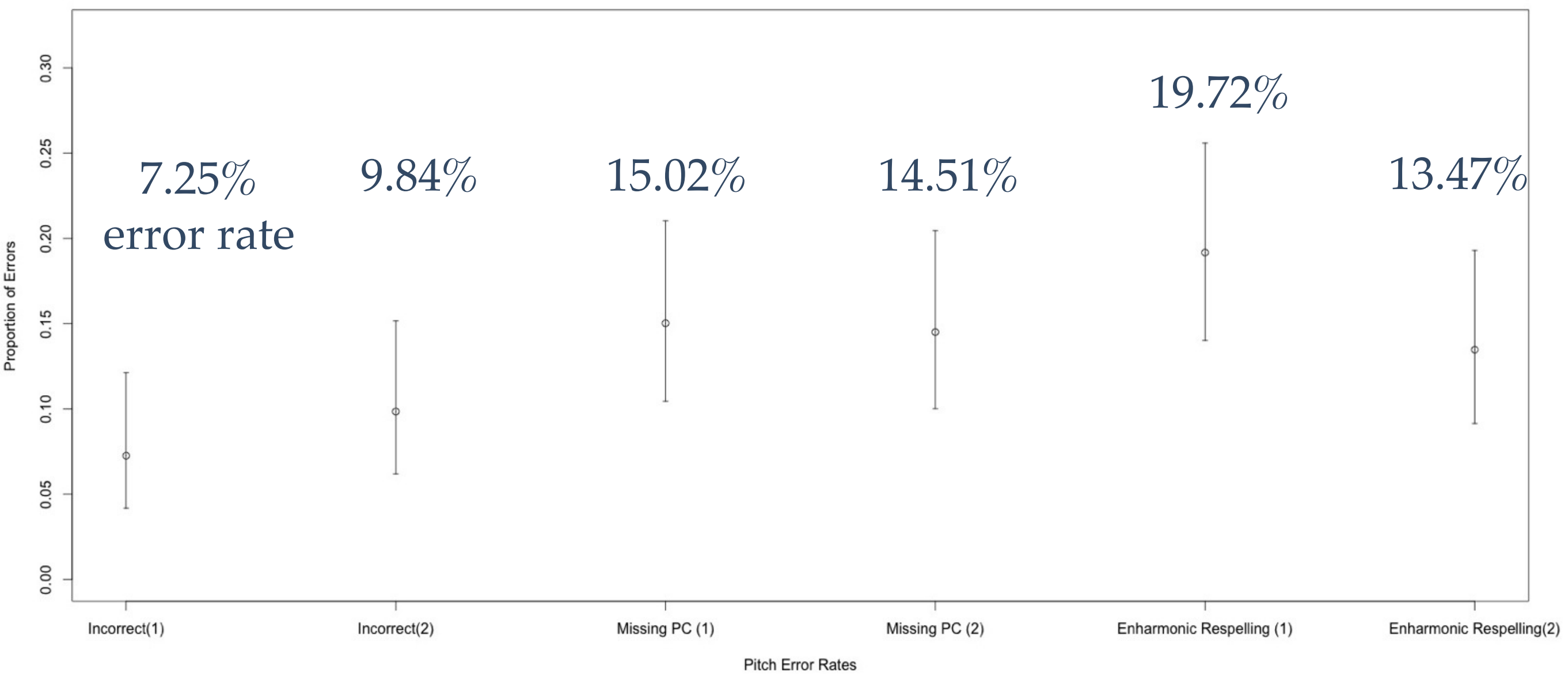




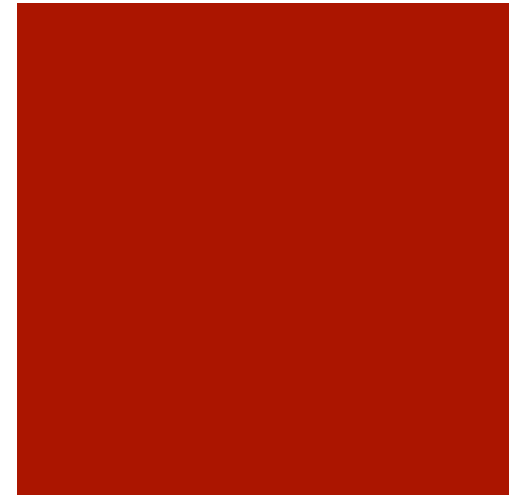
Encoding Error Rate vs. Analytic Error Rate



- “A single error in pitch encoding will affect the pitch pair comprising the interval, rendering the entire interval incorrect” (Huron, 1988).
 - Analytic Error = (Encoding Error) (Number of tokens in a single measurement)
 - Pitch errors examining large harmonic structures will have more of an effect than those found in smaller harmonic structures.
 - Similarly, one should adjust these error rates when analyzing n-grams of various lengths.



Rhythm Errors

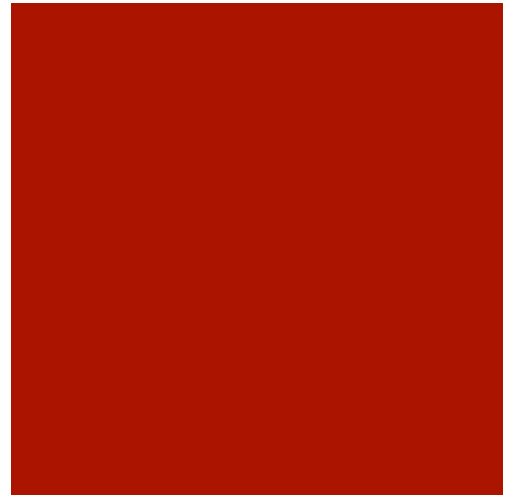


Onset position incorrect (Onset 1)

Onset position incorrect (Onset 2)

- If either onset 1 or onset 2 are on a different beat in the XML file than in the published score, count as an error.

Rhythm Errors

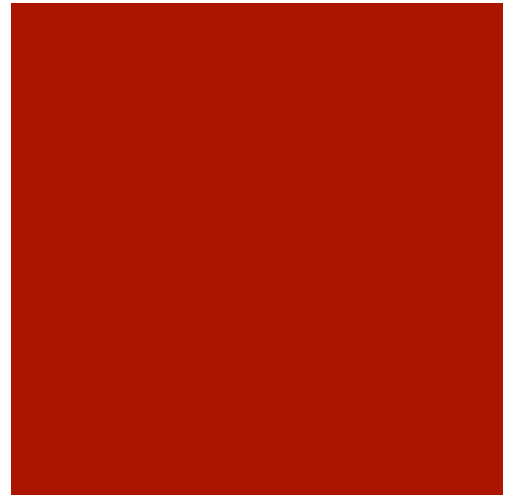


Durations incorrect (Onset 1)

Durations incorrect (Onset 2)

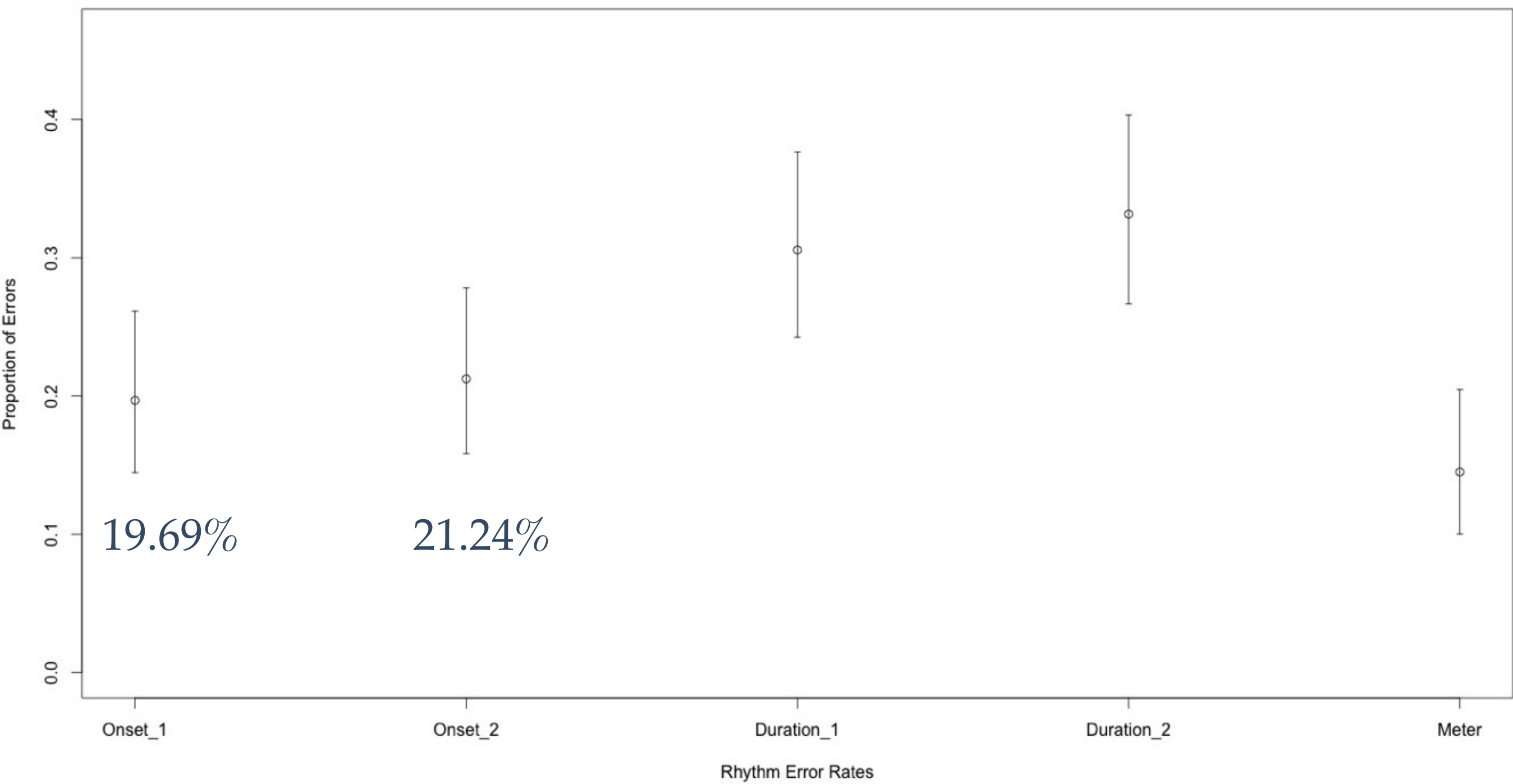
- If the onset is an incorrect duration (not counting the differences between ties and dots, etc.), count as 1 error. This applies to any of the voices in either onset 1 or 2.

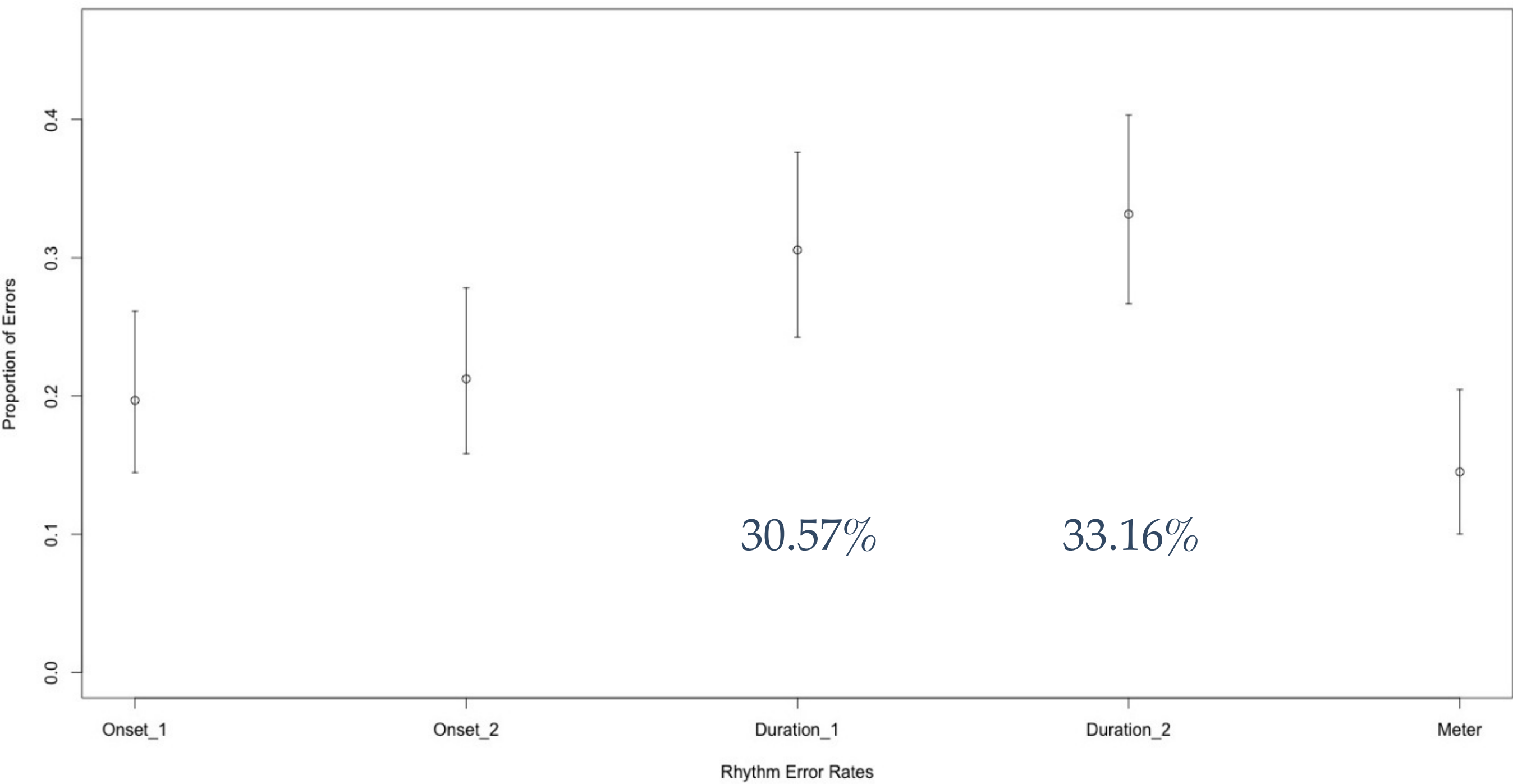
Rhythm Errors

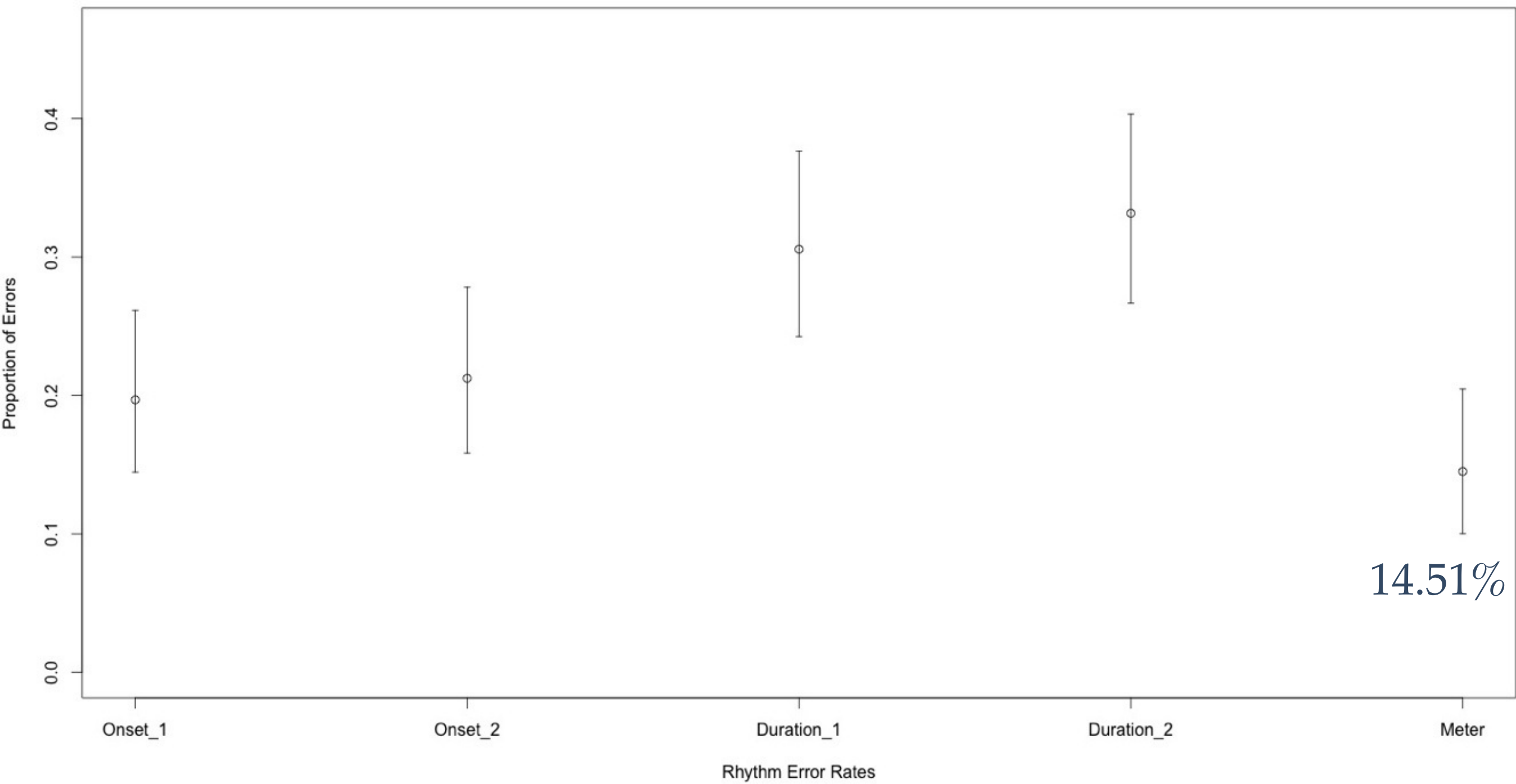


Meter is incorrect

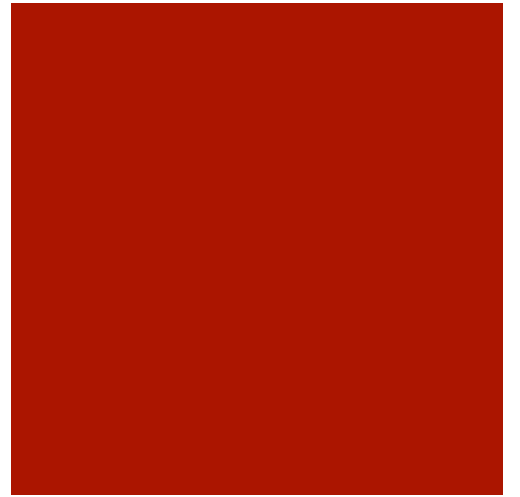
- If the notated meter is different from the notated meter *for any of the voices* count as 1 error. If some staves are correct, but others incorrect, use a fraction (we never experienced this).







Orchestration Errors



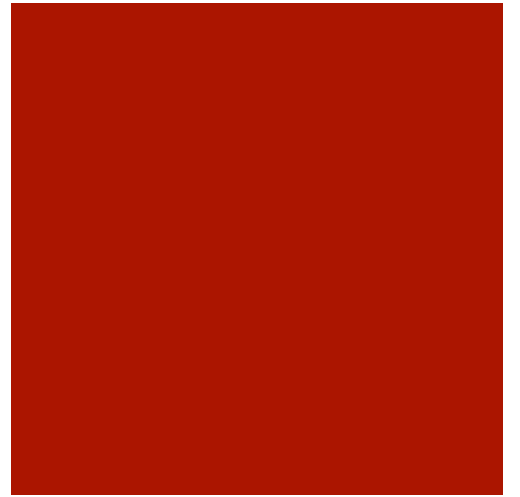
Missing doubling

- If a pitch is doubled across instruments in the published score, but not indicated as such in the XML, mark it as an error.

Extra doubling

- If a pitch is doubled across instruments in the XML, but not indicated as such in the published score, mark it as an error.

Orchestration Errors



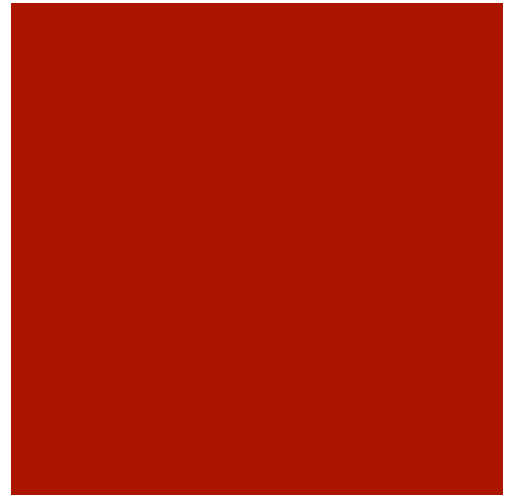
Note events misaligned across parts

- Consider the “mass of aligned onsets” to be the baseline. If an onset is not aligned, consider it an error.

Errors in instrument assignment

- If the word for the instrument in the score does not appear in the file in *any language*, count as 1 error. If an instrument in the file does not appear in the score, count as one error.

Orchestration Errors

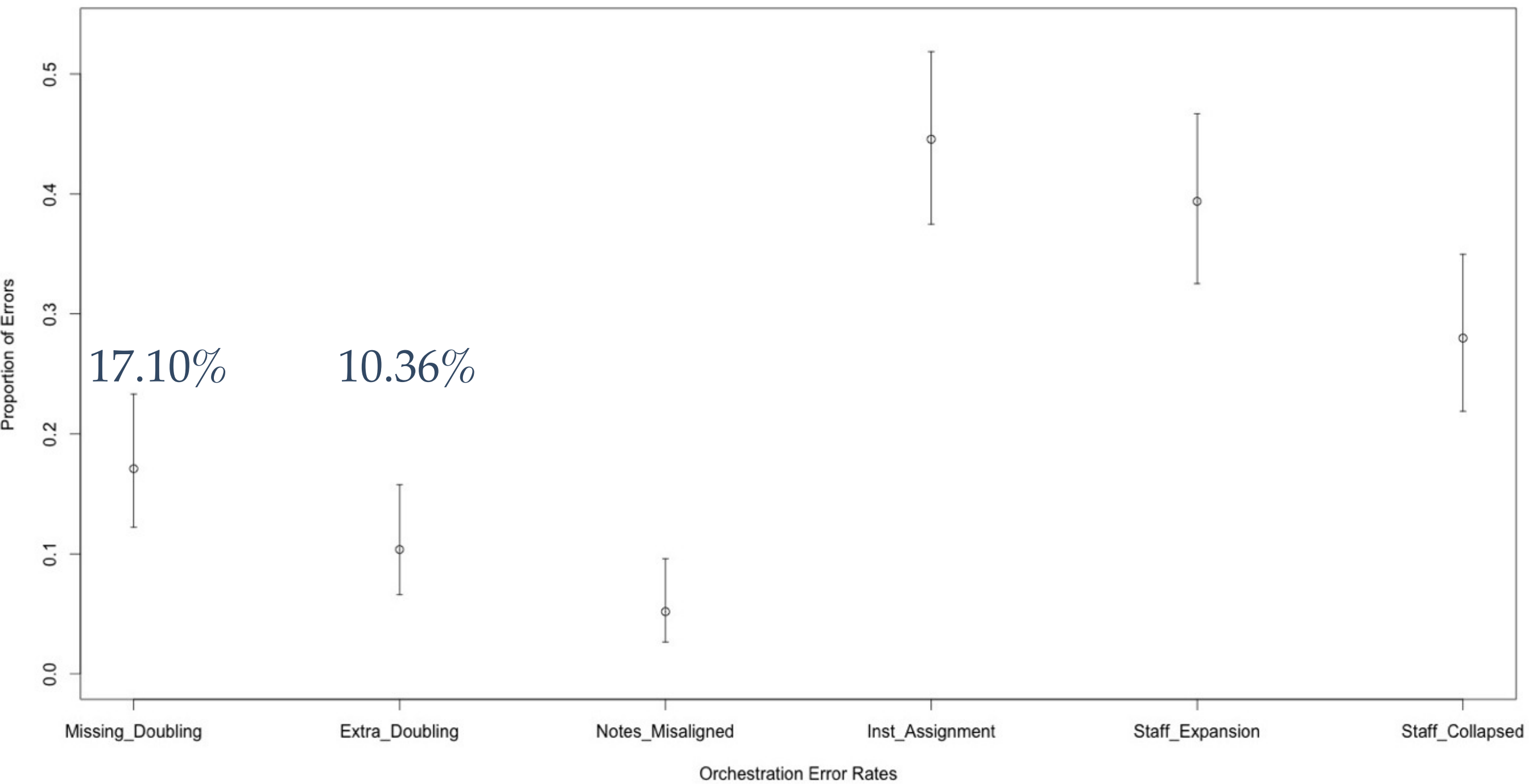


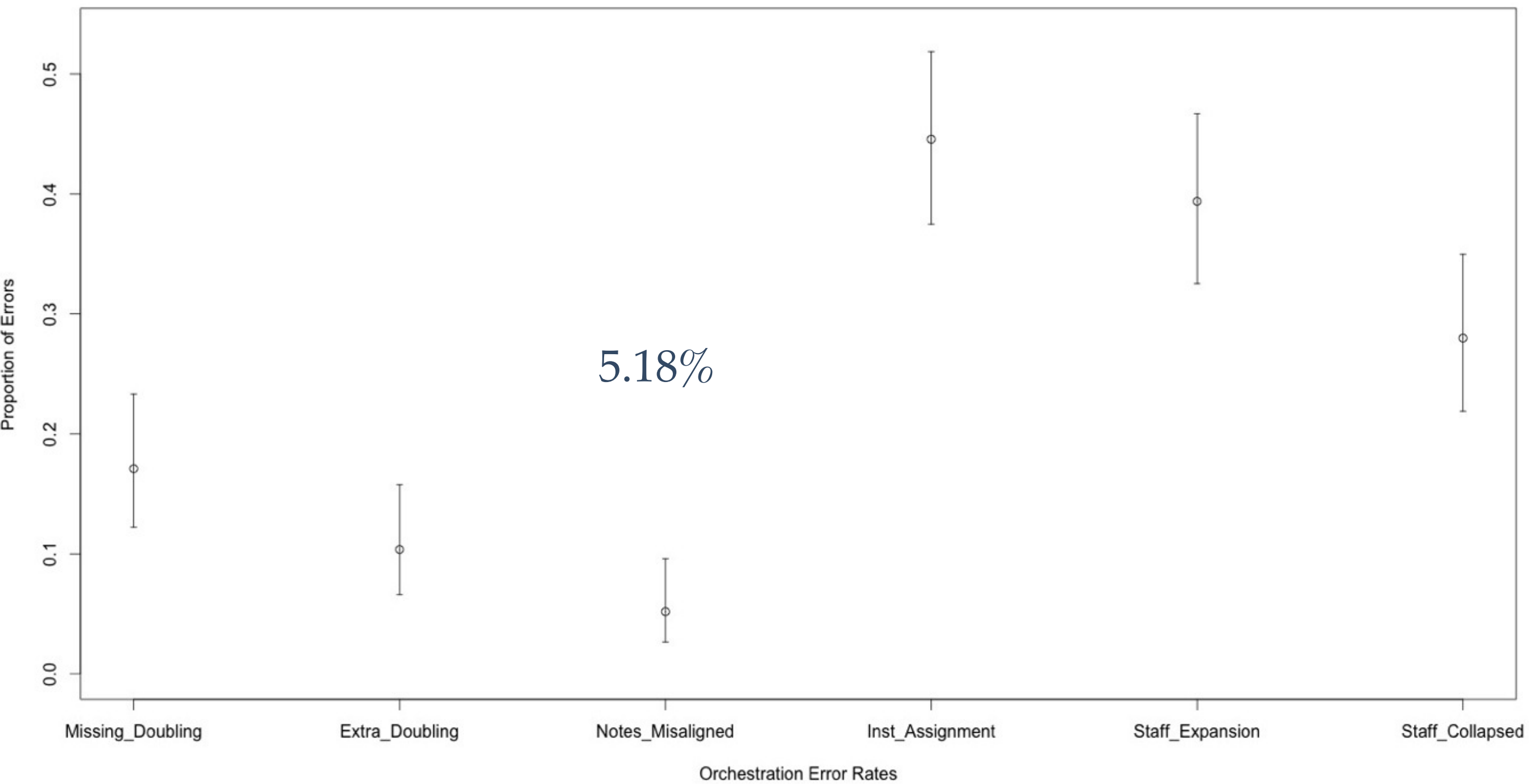
One staff spread across multiple staves

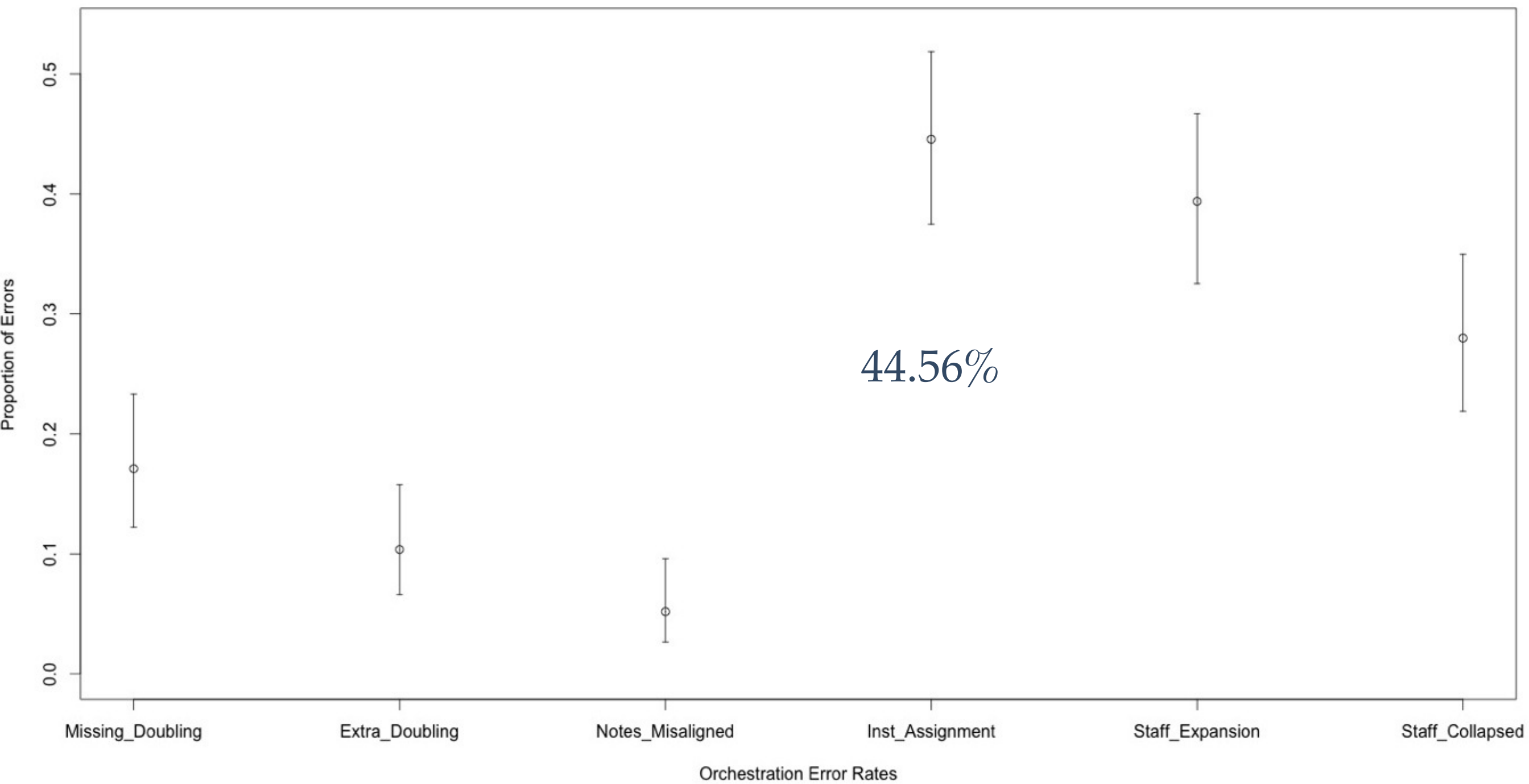
- If multiple voices on one staff are represented with an extra staff, it should be counted as one error.

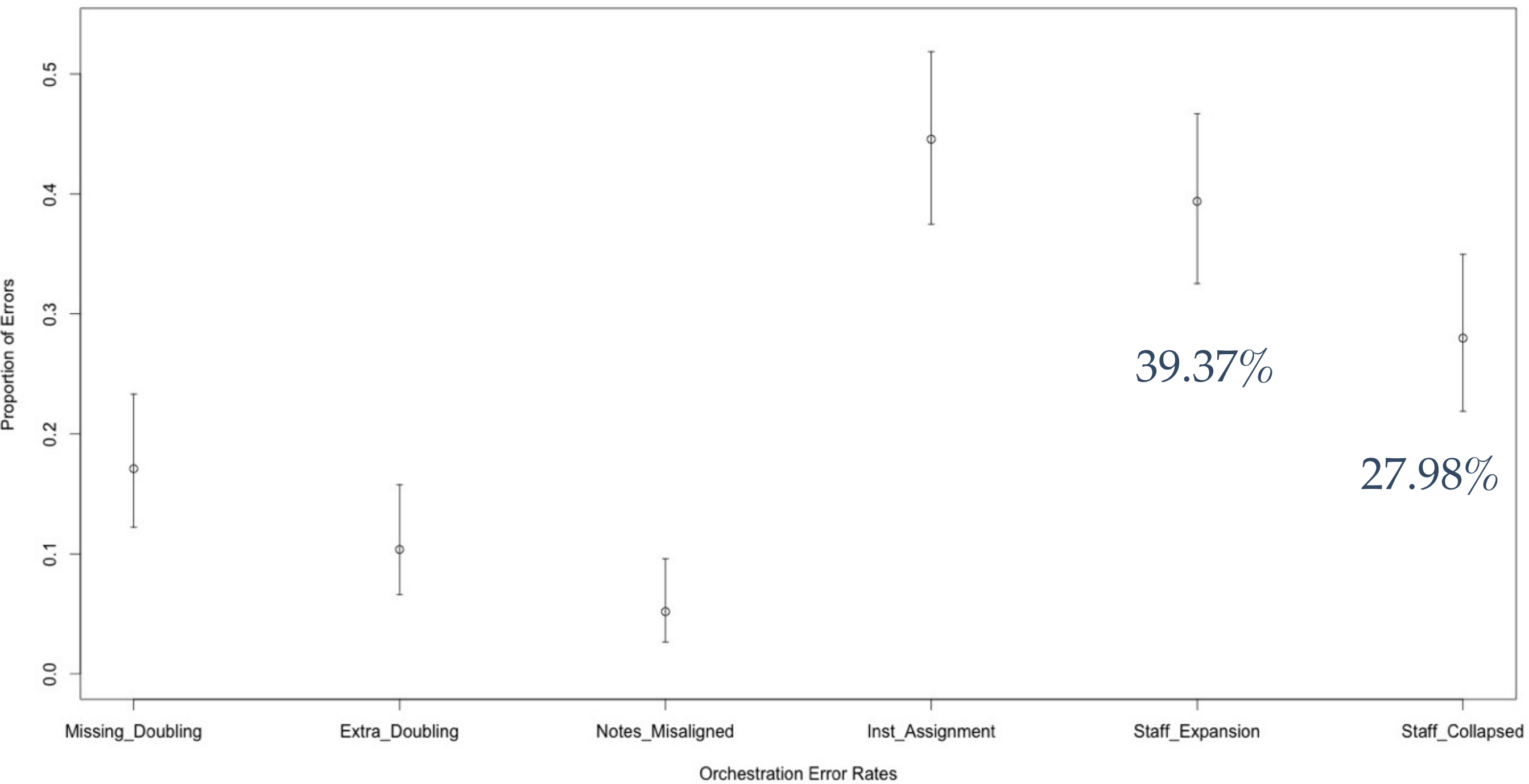
Multiple staves collapsed into one notated staff

- If staves are condensed into one staff or instrument, it should be counted as an error.

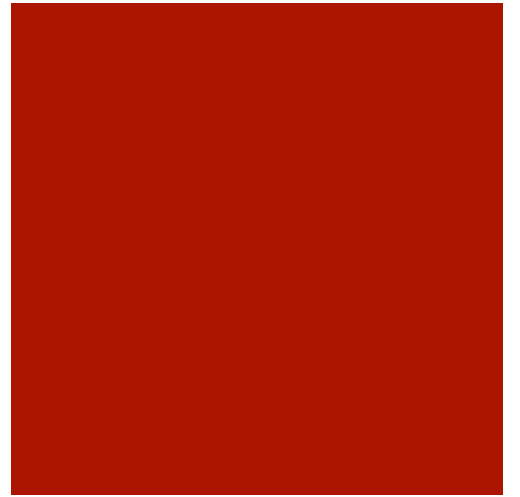








More Notation Errors



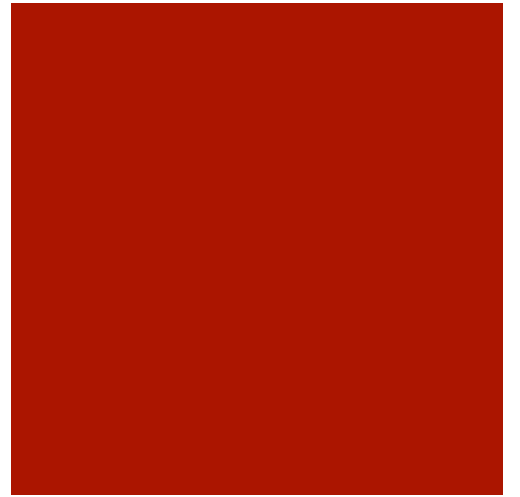
Discrepancy in notated clef

- If the notated clef in any of the instruments differs in the XML from the published score, mark it as an error.

Discrepancy in notated key signature

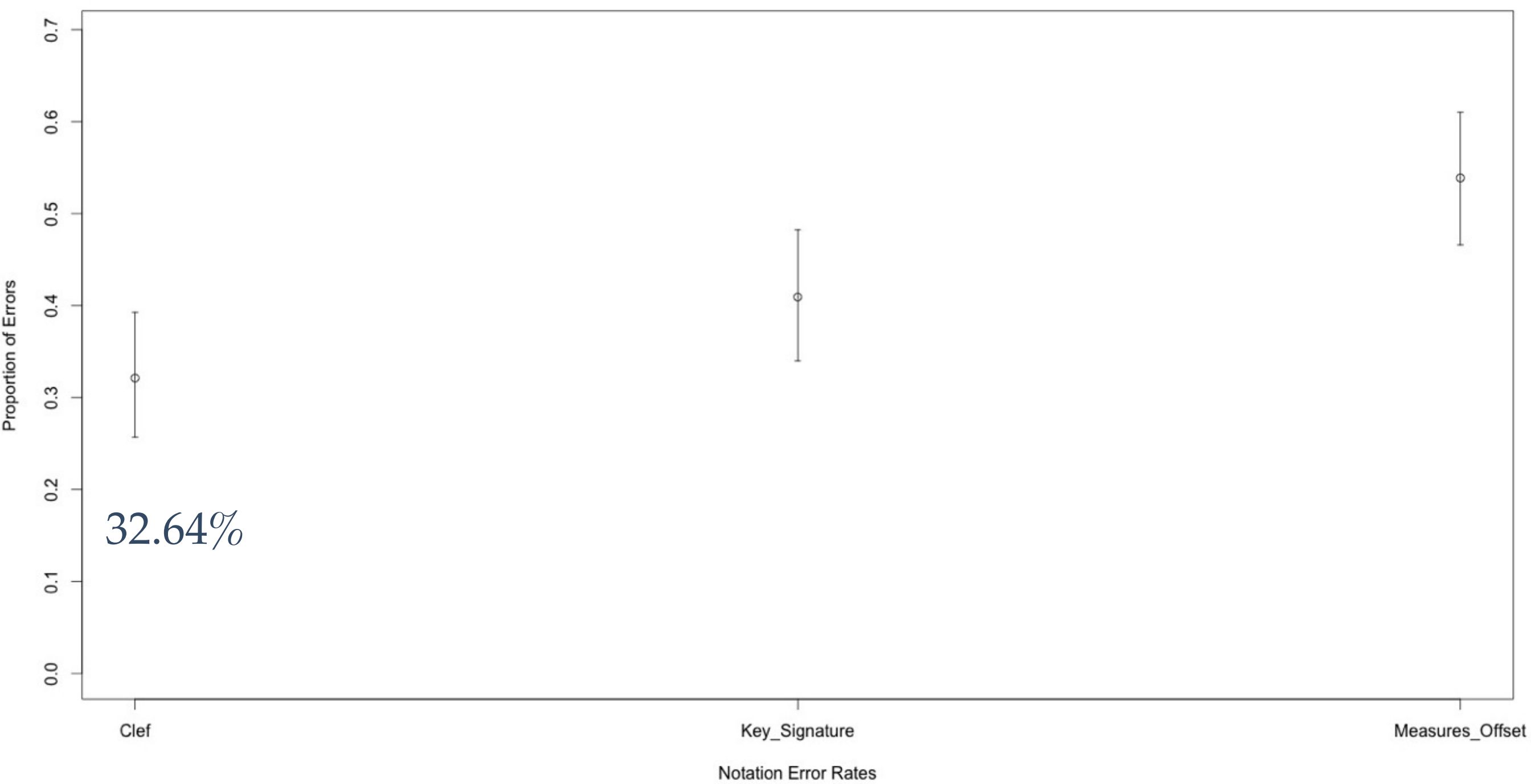
- If the key signature in any of the instruments differs in the XML from the published score, mark it as an error.

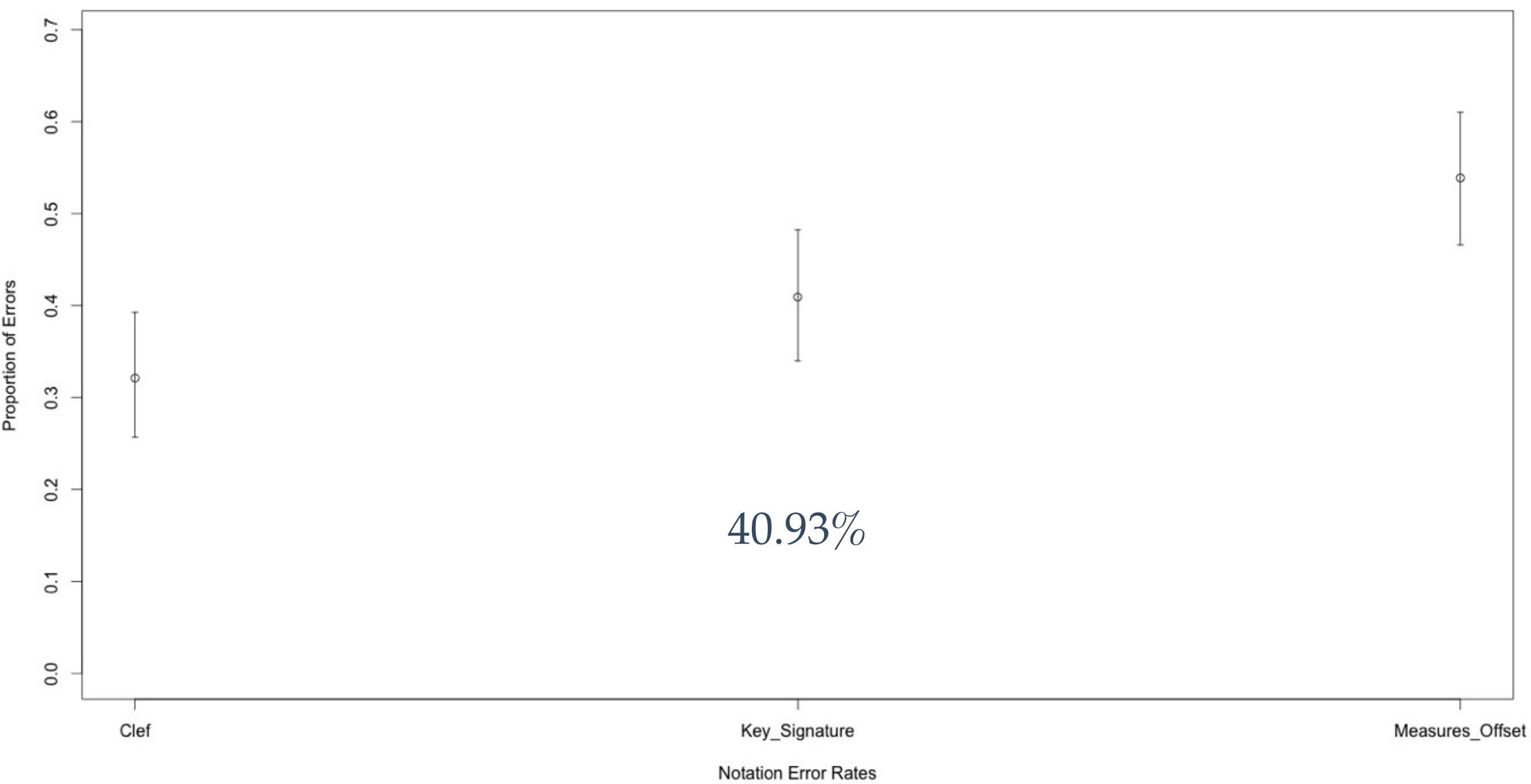
More Notation Errors

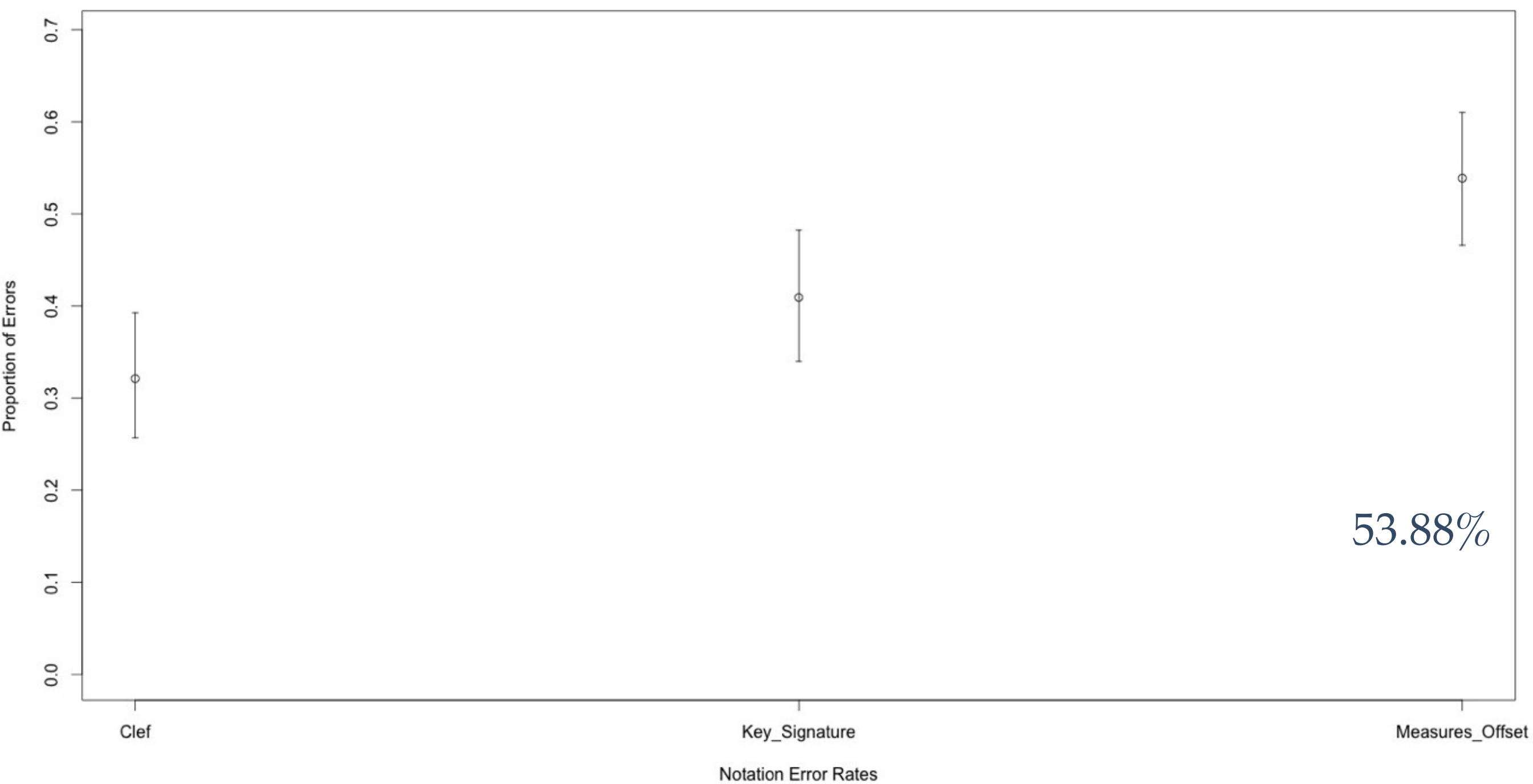


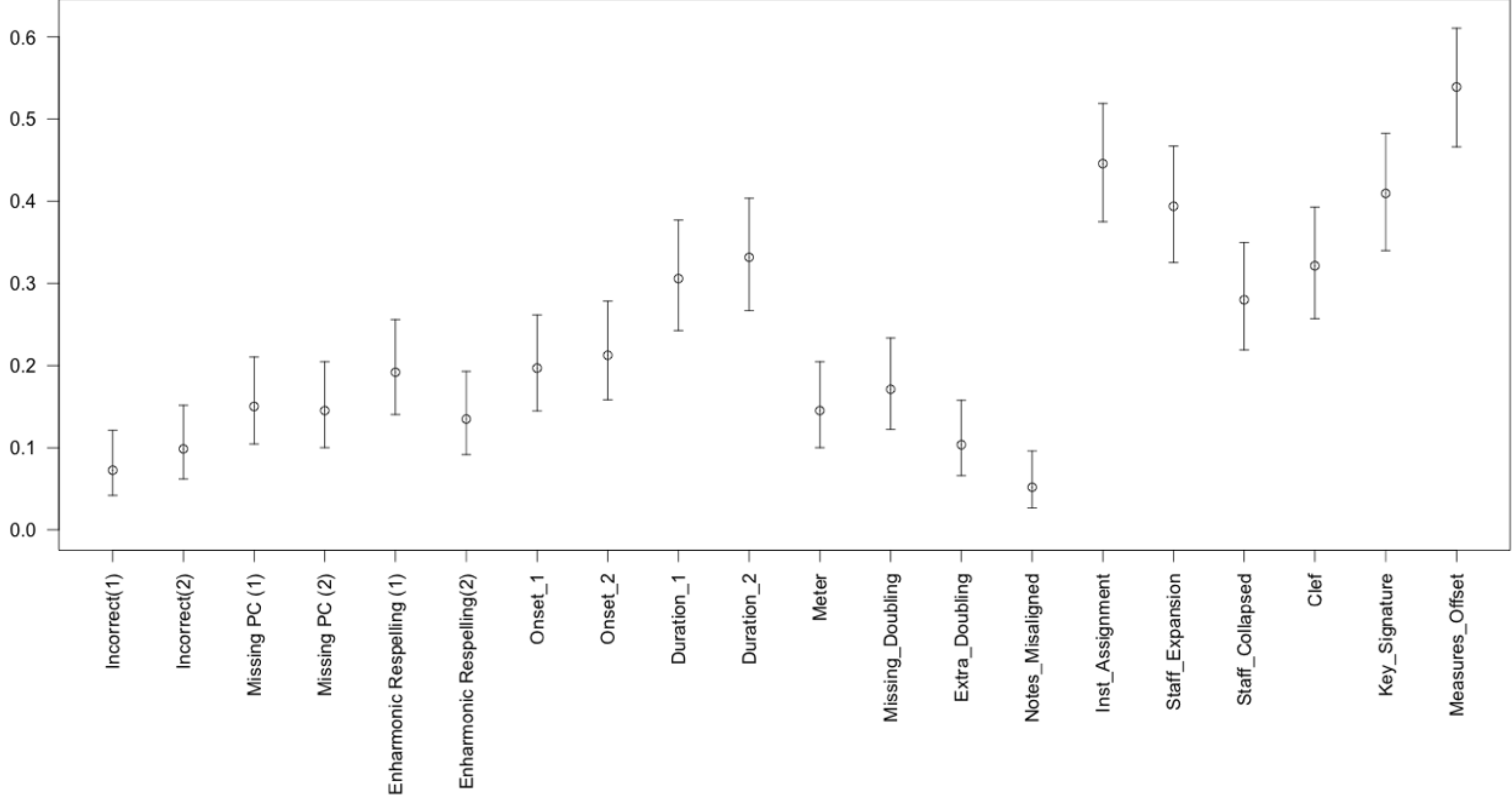
Offset measures in the beginning

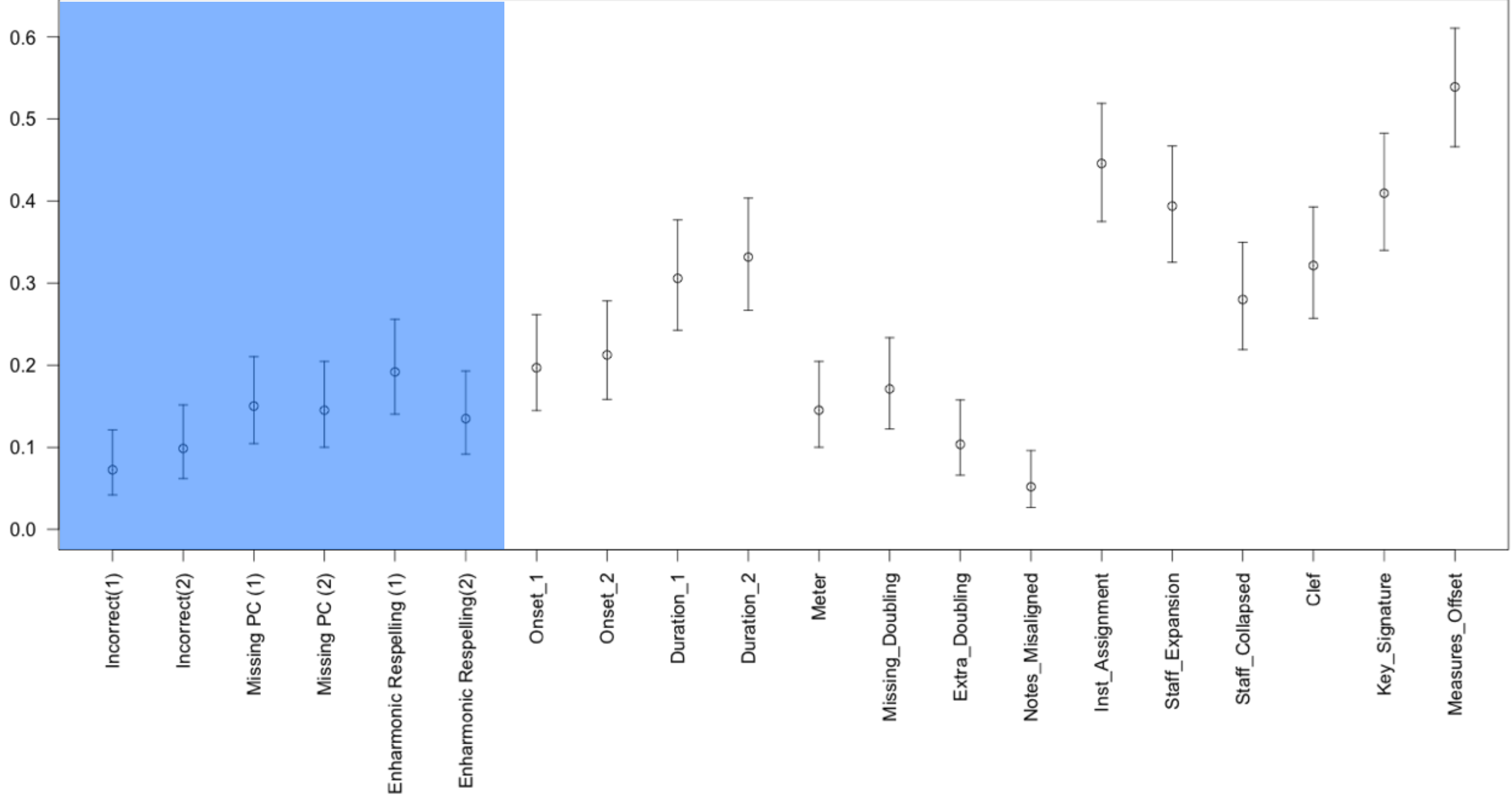
- If there are blank measures in the XML score before the beginning of the piece, mark it as an error.

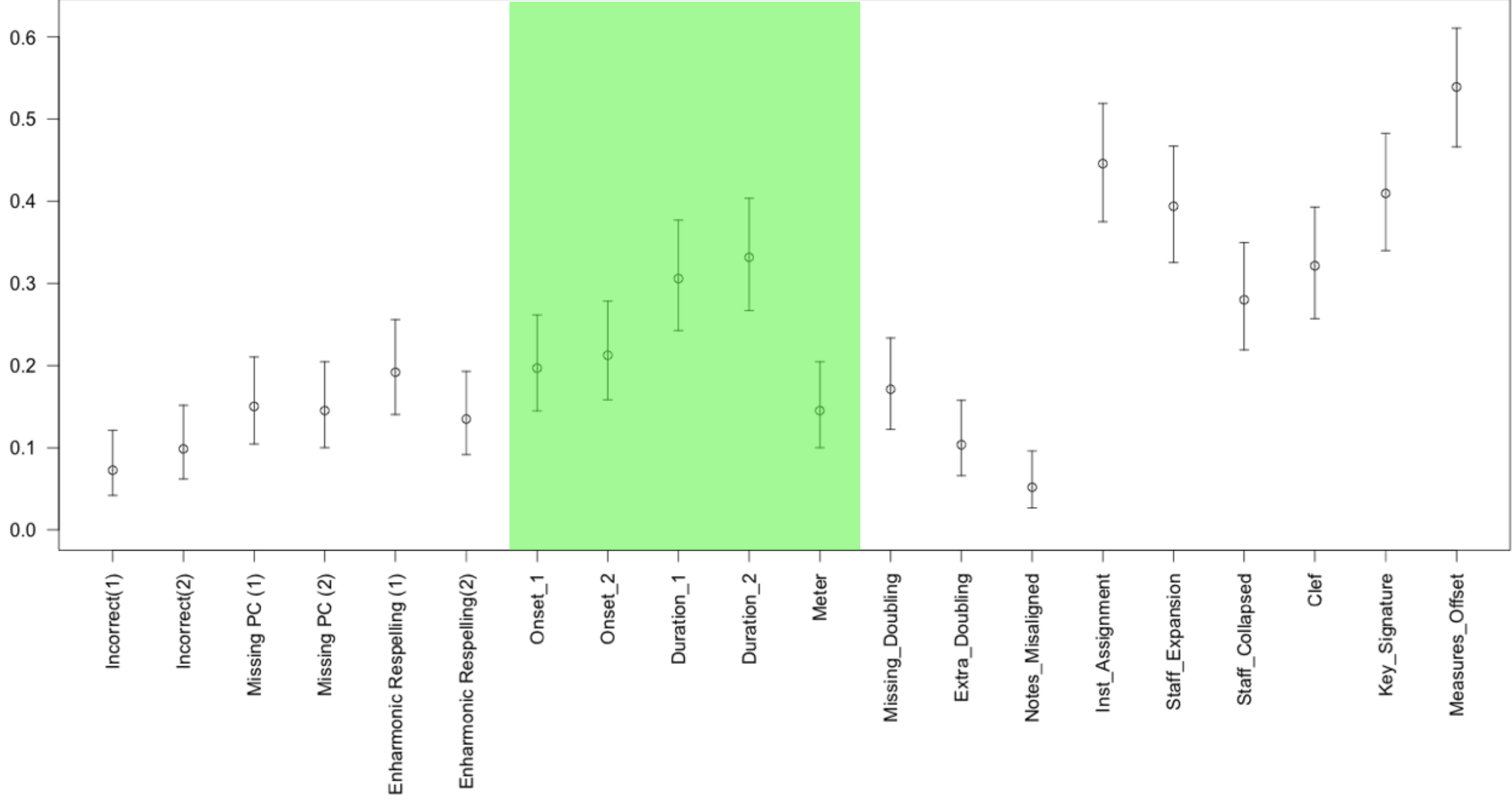


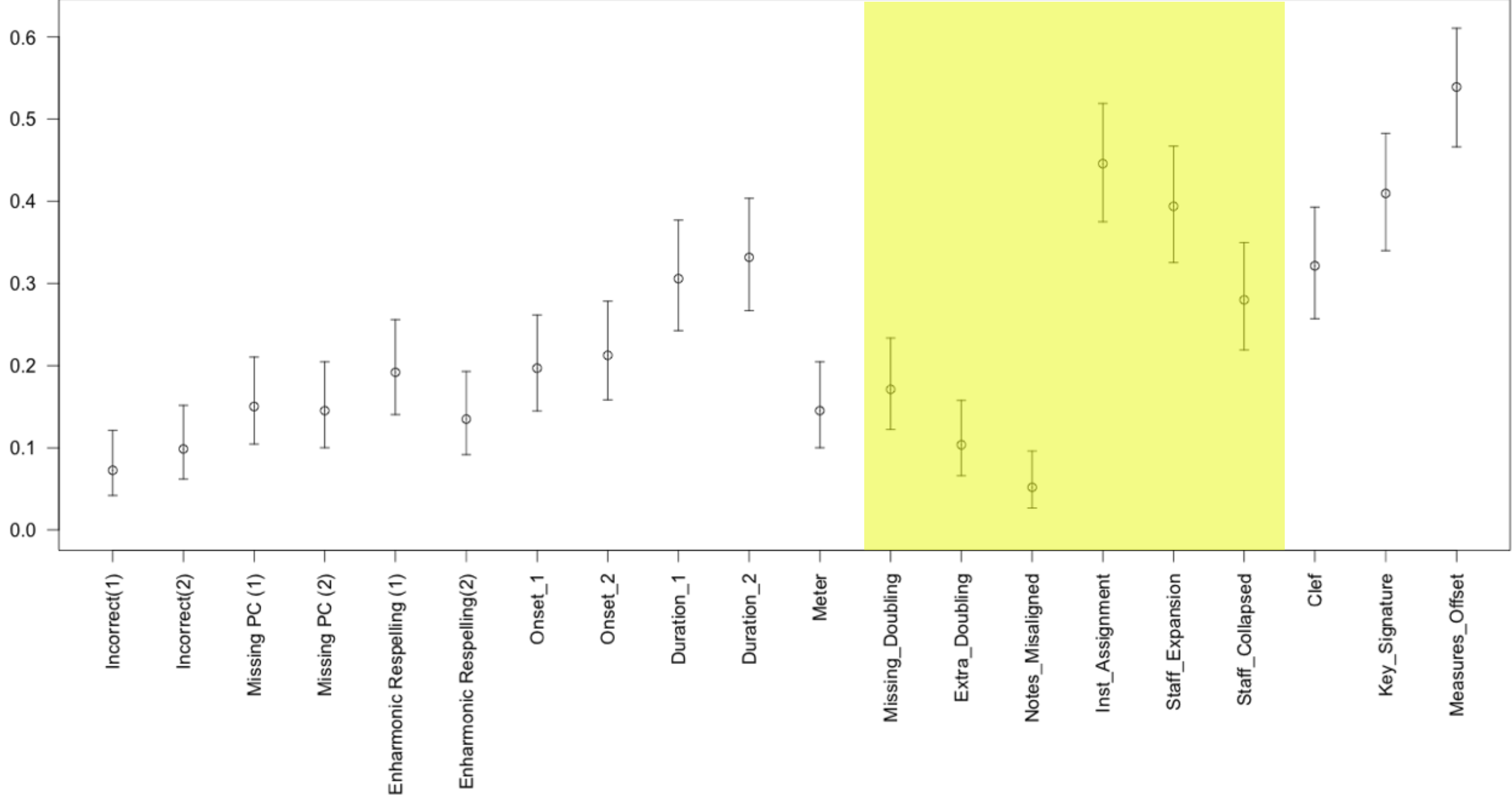


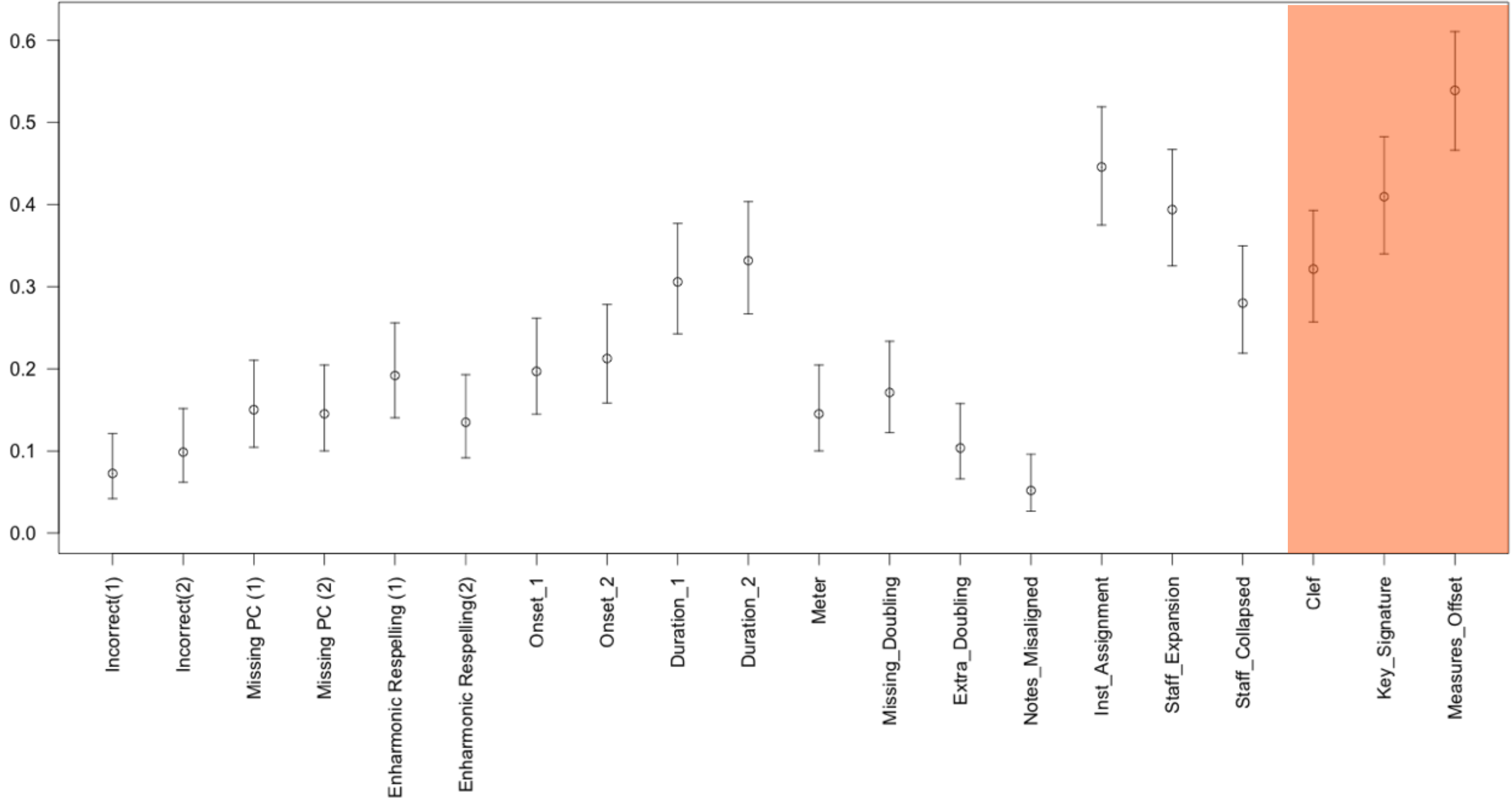


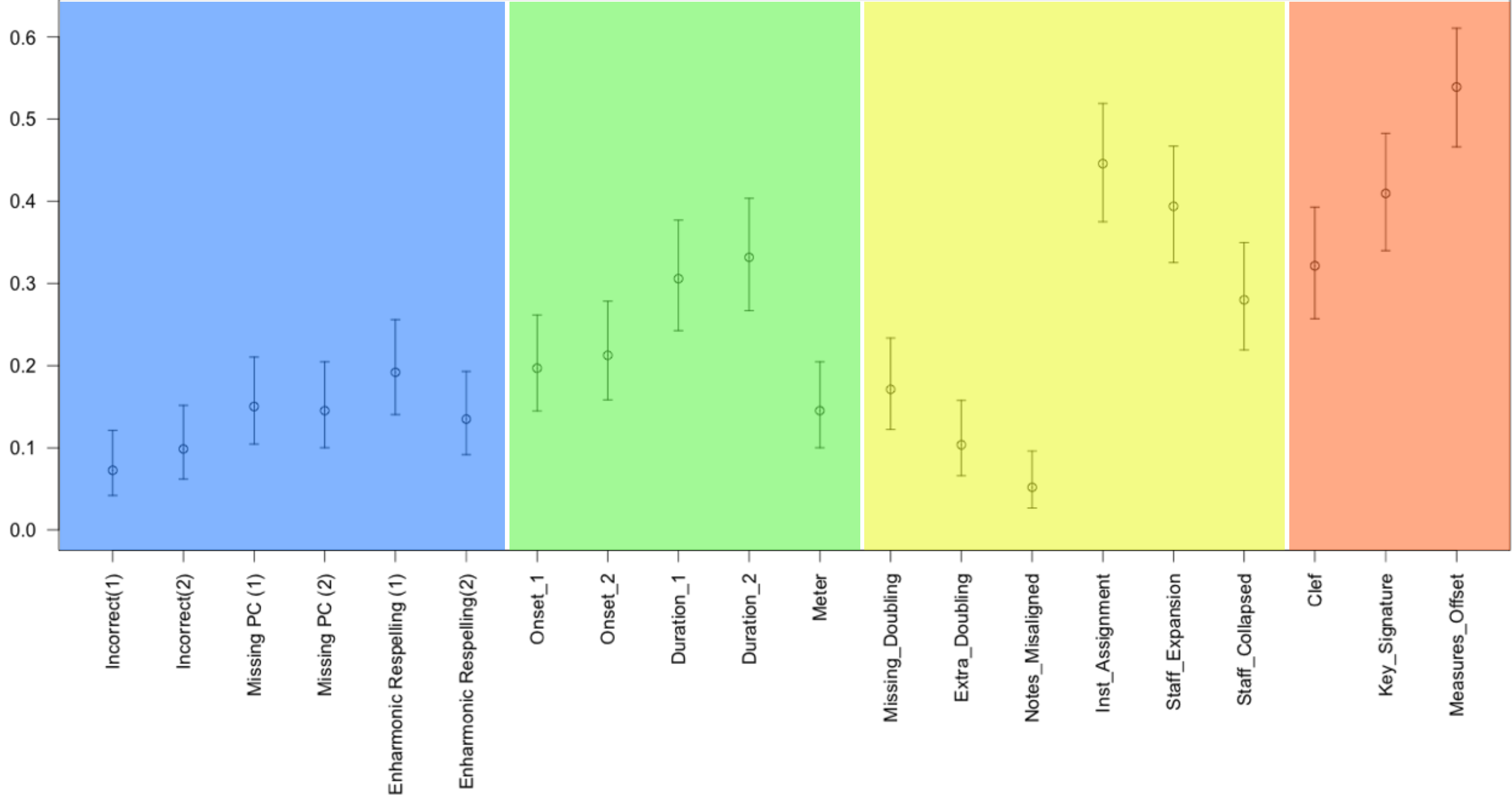












The Moral of the Story...



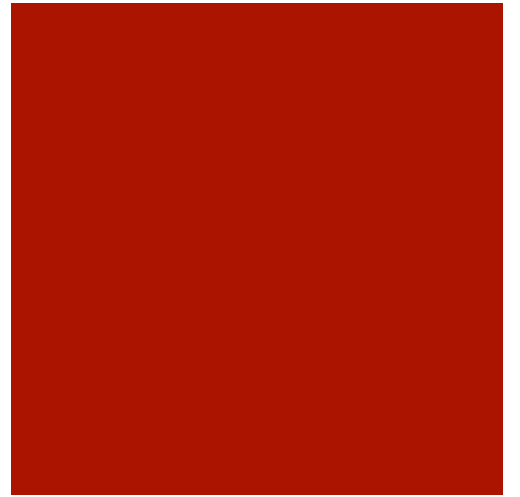
- It might seem that friends don't let friends use MIDI data.
 - However...
 - These errors weren't equally distributed. There was a very clear distinction between "good" and "bad" encodings.
 - If one were able to use only those which appeared to be manually encoded (which is often a visible distinction) the results would likely be much more accurate.



A MIDI Spam Filter?

Naïve Bayes Classifier

- 700 pieces were used to train a classifier.
- Both the MIDI data and the converted scores (in Finale) were both examined as whether or not they were likely mostly performed in or mostly step-input.



Problems

- By using IMSLP, it could be argued that we were nearly as reliant on canonical works as existing databases.
- IMSLP only allows scores out of copyright, so there are few 20th century composers represented.
- Similarly, “minor” composers are not as well represented as the “big names.”

Future Work

More validation

More cross-validation

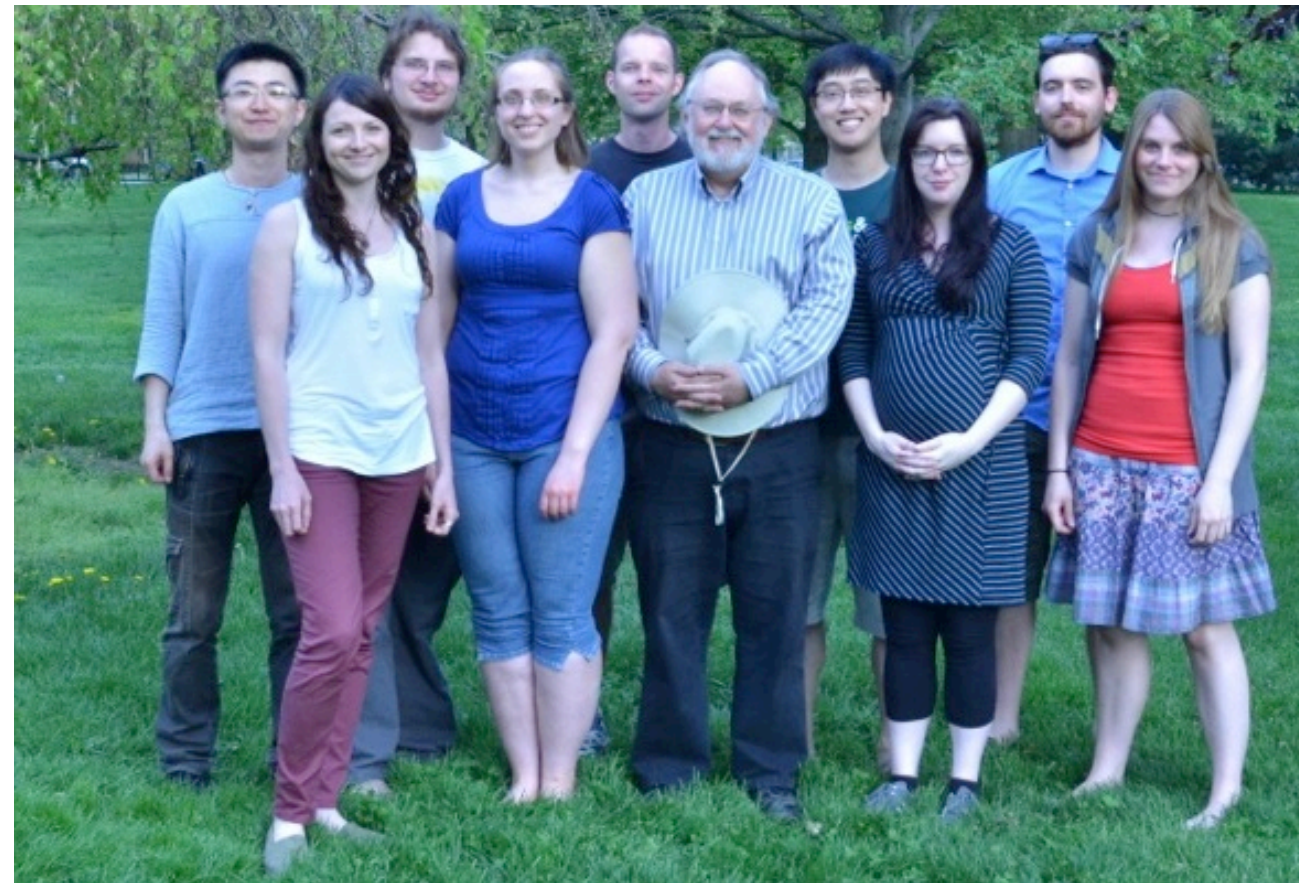
Future Work

More validation

More cross-validation

Cognitive and Systematic Musicology Lab, Ohio State University

Pierre Schwob, ClassicalArchives.com



References

Huron, David. "Error Categories, Detection, and Reduction in a Musical Database." *Computers and the Humanities* 22, no. 4 (January 1988): 253–264.