In [1]:

```
import pandas as pd
import numpy as np
import seaborn as sn
import datetime
import warnings
from operator import attrgetter
from tqdm import tqdm_notebook
from matplotlib import pyplot as plt
warnings.filterwarnings("ignore")

levels = ["Low to Minimal","Mild","Moderate","Severe"]
```

Data Preparation

My intuition for this analysis is to investigate how the degree of anxiety disorders can change temporally. So from the given database, I get time information in a month scale for now, which is number of months since first assessment. I also extract information such as score variation or level variation comparing with the score/level from first assessment.

Also, in order to assess the performance of therapies, **I assume the first assessment makes the beginning of the treatment**, and I look into how the treatment works along the time in comparison with the results form first assessment.

In [2]:

```
data = pd.read_csv("phq_all_final.csv")

data['date'] = pd.to_datetime(data['date'], format='%Y-%m-%dT%H:%M:%S.
%f')
data['patient_date_created'] = pd.to_datetime(data['patient_date_creat
ed'], format='%Y-%m-%dT%H:%M:%S.%f')
data['current_month_year'] = pd.to_datetime(data['date']).dt.to_period
('M')

data=data.sort_values(by=['date']).reset_index()

# labels of a patient's level of anxiety disorders
bins = [-1,6,11,16,22]
labels = [0,1,2,3]
data['level']=pd.cut(data.score,bins,False,labels).astype('int64')

# labels of a patient is in a severe condition or not
bins_severe = [-1,10,22]
labels_severe = [0,1]
data['severe']=pd.cut(data.score,bins_severe,False,labels_severe).asty
pe('int64')
```

```
patients = data['patient_id'].unique()
# create logs for each patient
patient_logs = {}
for patient in tqdm_notebook(patients):
    patient_log = data[data['patient_id']==patient]

    first_score = patient_log.iloc[0]['score']
    first_level = patient_log.iloc[0]['level']
    first_severity = patient_log.iloc[0]['severe']
    first_date = patient_log.iloc[0]['date']
    first_month_year = patient_log.iloc[0]['current_month_year']

    # mark the metrics from first assessment
    data.loc[data.patient_id==patient,'first_score']=first_score
    data.loc[data.patient_id==patient,'first_level']=first_level
    data.loc[data.patient_id==patient,'first_date']=first_date
    data.loc[data.patient_id==patient,'first_month_year']=first_month_year

    patient_log = data[data['patient_id']==patient]
    patient_logs[patient] = patient_log

# temporal distance
data['month_distance'] = (
    data['current_month_year'] -
    data['first_month_year']).apply(attrgetter('n'))

# variation in assessment results
data['score_variation'] = (
    data['score'] -
    data['first_score'])

data['level_variation'] = (
    data['level'] -
    data['first_level'])

data.drop(columns=['index'])
```

Out[2]:

```
                date  patient_id  type  patient_date_created  score  current_month_year  leve

    0     2019-06-06        9834  gad7            2019-06-04     15             2019-06
     16:31:34.960999                   13:20:52.492773

    1     2019-06-07        9130  gad7            2019-06-03      0             2019-06
     07:17:53.394337                   21:14:27.617500

    2     2019-06-07        3788  gad7            2019-06-03      0             2019-06
     13:05:01.435941                   14:48:09.129756

    3     2019-06-09        3915  gad7            2019-06-04      4             2019-06
     16:44:23.212699                   19:22:24.754240

    4     2019-06-09       16617  gad7            2019-06-04     12             2019-06
     22:54:50.120132                   19:28:38.734048

  ...            ...         ...   ...                   ...    ...                 ...

53693     2020-07-31        9662  gad7            2020-04-06      7             2020-07
     19:53:45.988090                   22:04:28.746347

53694     2020-07-31         999  gad7            2020-02-04      3             2020-07
     20:01:26.115163                   21:40:55.624196

53695     2020-07-31       12696  gad7            2020-05-05      4             2020-07
     20:01:56.773677                   19:50:04.906831

53696     2020-07-31        2912  gad7            2020-03-23      4             2020-07
     20:05:53.520605                   16:08:39.976328

53697     2020-07-31       12589  gad7            2020-04-28      5             2020-07
     20:11:33.663924                   16:23:07.483774
```

53698 rows × 15 columns

In [3]:

```
data.head()
```

Out[3]:

```
   index             date  patient_id  type  patient_date_created  score  current_month_year  l

0     58       2019-06-06        9834  gad7            2019-06-04     15             2019-06
          16:31:34.960999                   13:20:52.492773

1     52       2019-06-07        9130  gad7            2019-06-03      0             2019-06
          07:17:53.394337                   21:14:27.617500

2      7       2019-06-07        3788  gad7            2019-06-03      0             2019-06
          13:05:01.435941                   14:48:09.129756

3    132       2019-06-09        3915  gad7            2019-06-04      4             2019-06
          16:44:23.212699                   19:22:24.754240

4    140       2019-06-09       16617  gad7            2019-06-04     12             2019-06
          22:54:50.120132                   19:28:38.734048
```

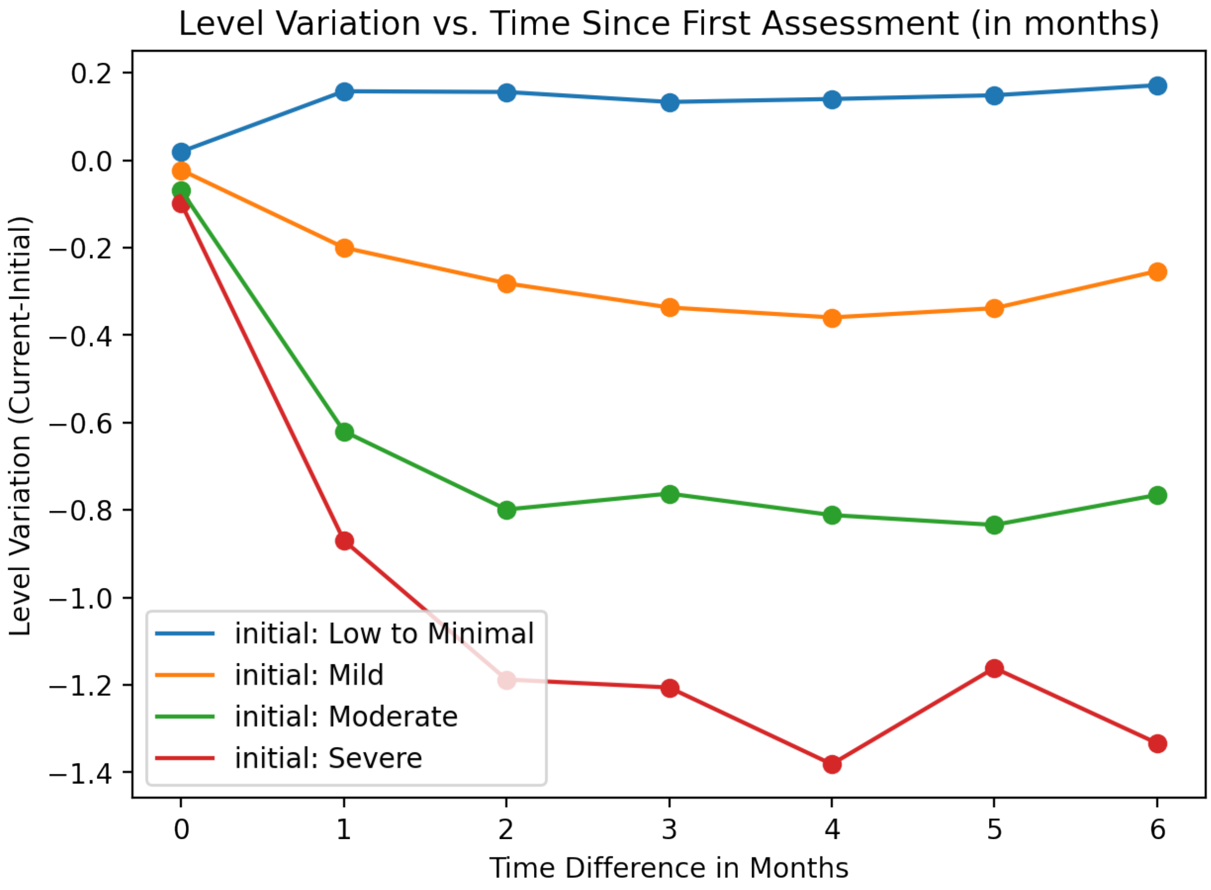
Data Analysis

In [4]:
```
data_analysis = data[data['date']!=data['first_date']] # remove the data on first date
```

In this following plot, I analyzed the relationship between the change in axiety levels and the months since first assessment (treatment). The visualization shows that apart from the initial level at Low to Minimal, on average, the treatment shows consistent reduction on anxiety levels for the other three initial levels, i.e, considerable initial anxiety levels can be reduced and maintained at a lower level by treatment throughout the therapies. **This shows the effectiveness of the treatment**.

However, for patients start with negligible anxiety disorders, it may be more reasonable to offer them different treatment than the other patients. The plot shows over-treatment can lead to considerable increase in a patient's anxiety level if he/she starts with little anxiety. **Other measures to help them maintain the current status would be a better way**.

In [5]:
```
fig, ax1 = plt.subplots(figsize=(7,5), dpi=200)
variation_by_first = data_analysis.groupby(['first_level','month_distance'])['level_variation'].mean()
for i in range(4):
    plt.plot(variation_by_first[i].index,variation_by_first[i].values,label='initial: {}'.format(levels[i]))
    plt.scatter(variation_by_first[i].index,variation_by_first[i].values)
plt.xlabel("Time Difference in Months")
plt.ylabel("Level Variation (Current-Initial)")
plt.title("Level Variation vs. Time Since First Assessment (in months)")
plt.legend()
plt.show()
```

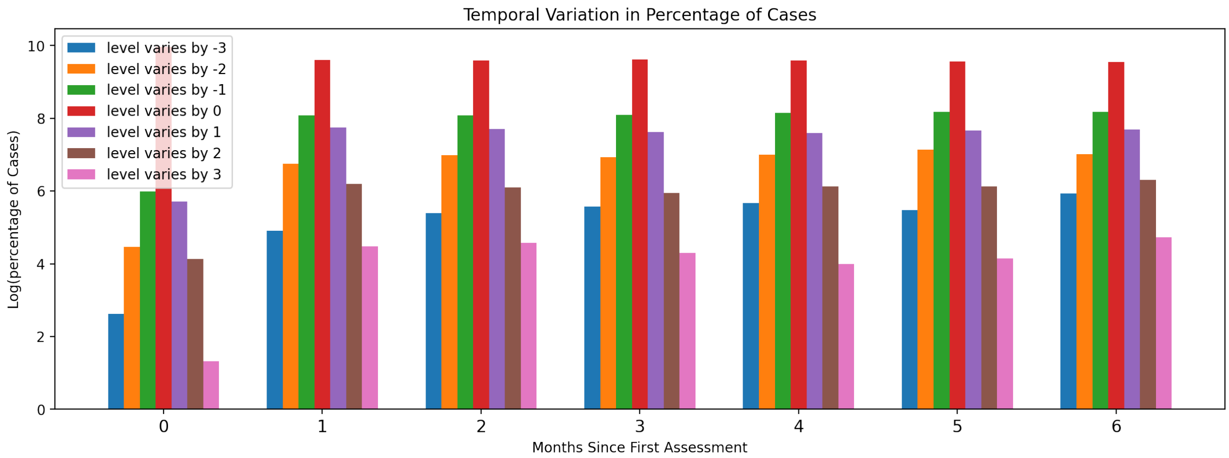


The following section shows how different values of anxiety level change take up the entire set of variation cases. We can see the log-scaled percentage of level variation of -1, -2, and -3 (anxiety level decreases) turn stable in just one to two month. The possibility can be that therapies take effect in one to two months, and then tend to hold its effect in the future. Or, according to the histogram shown later, it's very possible that **many people drop the sessions after they take their first assessment in the first month and feel no difference**, and the remaining population mostly have the same state in the following months, which leads to the stable distribution.

Therefore, if the second conclusion holds, it's important to look into why people tend to drop out after one month, through surveys or medical records, there can be certain features that can affect people's decision on whether to drop out or not. Based on the apparent decrease of cases where patients' anxiety level variation equals 0, **how to keep people with negligible anxiety disorders in the sessions** is an important topic to look into.

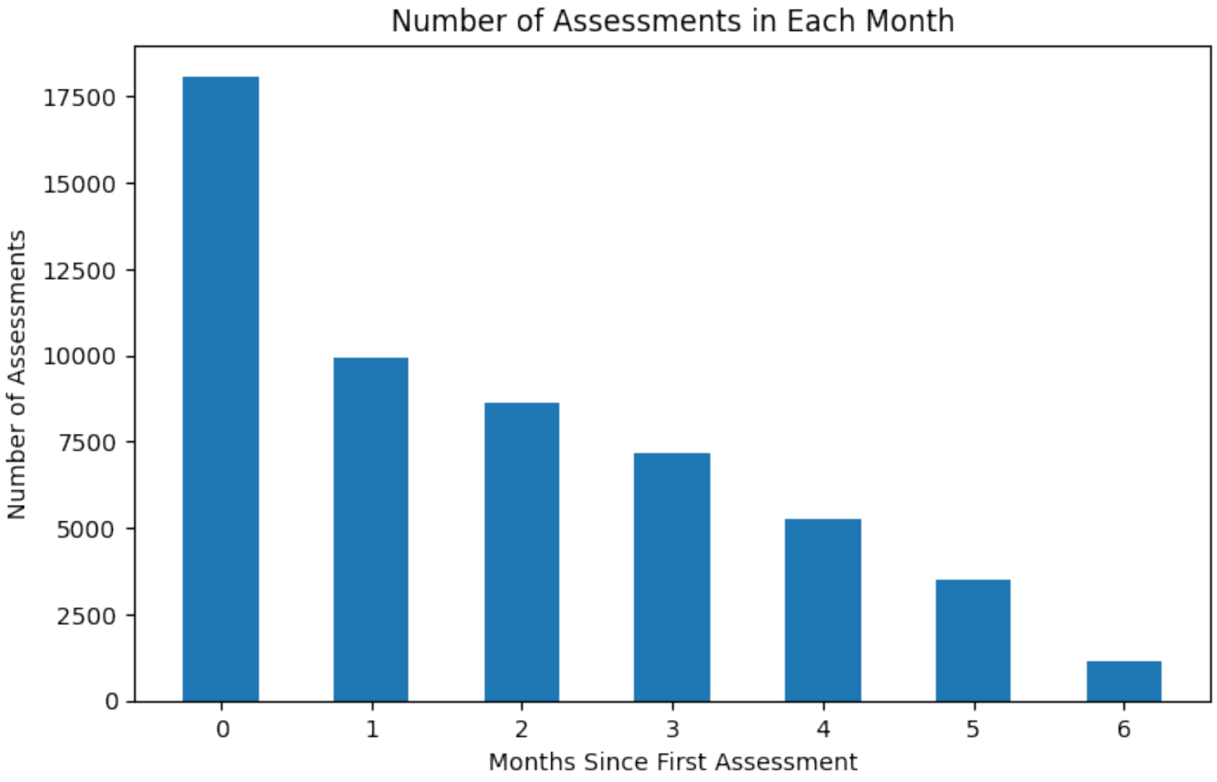
In [6]:

```python
fig, ax1 = plt.subplots(figsize=(15,5), dpi=200)
num_variation_distance=data_analysis.groupby(['level_variation','month_distance'])['patient_id'].count()
monthly_count = data_analysis.groupby(['month_distance'])['patient_id'].count()
width = 0.1
for i in range(-3,4):
    x = num_variation_distance[i].index
    y = num_variation_distance[i].values
    y = np.log([y[i]/monthly_count[i] for i in range(7)])+10
    plt.bar(x+width*i, y, width, label='level varies by {}'.format(i))
plt.xticks(x, fontsize=12)
plt.xticks(x, x)
plt.xlabel("Months Since First Assessment")
plt.ylabel("Log(percentage of Cases)")
plt.title("Temporal Variation in Percentage of Cases")
plt.legend()
plt.show()
```



In [7]:

```
fig, ax1 = plt.subplots(figsize=(8,5), dpi=100)
count = data.groupby('month_distance')['patient_id'].count()
plt.bar(count.index, count.values, 0.5)
plt.xlabel("Months Since First Assessment")
plt.ylabel("Number of Assessments")
plt.title("Number of Assessments in Each Month")
plt.show()
```



The recommended threshold for further clinical evaluation is also a variable I looked into. I inverstigated the percentage of patients that turn their severe status in the beginning around to a score under 10, as well as who become a severe one from a non-severe one, and how many times they take an assessment. Interestingly, the statistics shows **no significant difference in the average number of days they participate in the assessment**. This can be an indicator that the outcome of therapies is not correlated with the time a patient spends but rather other features such as the process of treatment. Here I assume one time of assessment as a treatment again.

In [8]:

```
reverse_bad = []
reverse_bad_times = []
reverse_good = []
reverse_good_times = []
for patient in tqdm_notebook(patients):
    patient_log = data[data['patient_id']==patient]

    # people who gain their anxiety disorders from initially non-severe level
    if patient_log.iloc[0]['severe']==0:
        result = patient_log.iloc[-1]['severe']==1
        if result:
            reverse_bad_times.append(len(patient_log))
        reverse_bad.append(result)
    # people who lower their anxiety disorders from initially severe level
    if patient_log.iloc[0]['severe']==1:
        result = patient_log.iloc[-1]['severe']==0
        if result:
            reverse_good_times.append(len(patient_log))
        reverse_good.append(result)
```

In [9]:

```
good_rate = sum(reverse_good)/len(reverse_good)
bad_rate = sum(reverse_bad)/len(reverse_bad)
print("%0.2f%% initially severe patients has lower their socres to under the threshold, average assessment times:%0.2f"%(good_rate*100, np.mean(reverse_good_times)))
print("%0.2f%% initially non-severe patients has gained anxiety level to above the threshold, average assessment times:%0.2f"%(bad_rate*100, np.mean(reverse_bad_times)))
```

```
31.55% initially severe patients has lower their socres to under the threshold, average assessment times:4.97
4.96% initially non-severe patients has gained anxiety level to above the threshold, average assessment times:4.20
```

In the future of this analysis, **finer-grained data analysis** should be conducted to find more detailed pictrues about how people's anxiety level varies throughout the time, maybe on a weekly or daily basis to analyze the functions of different kinds of treatment, to see how they create different time series data and find out the most effective treatment arrangement with respect to time intervals.

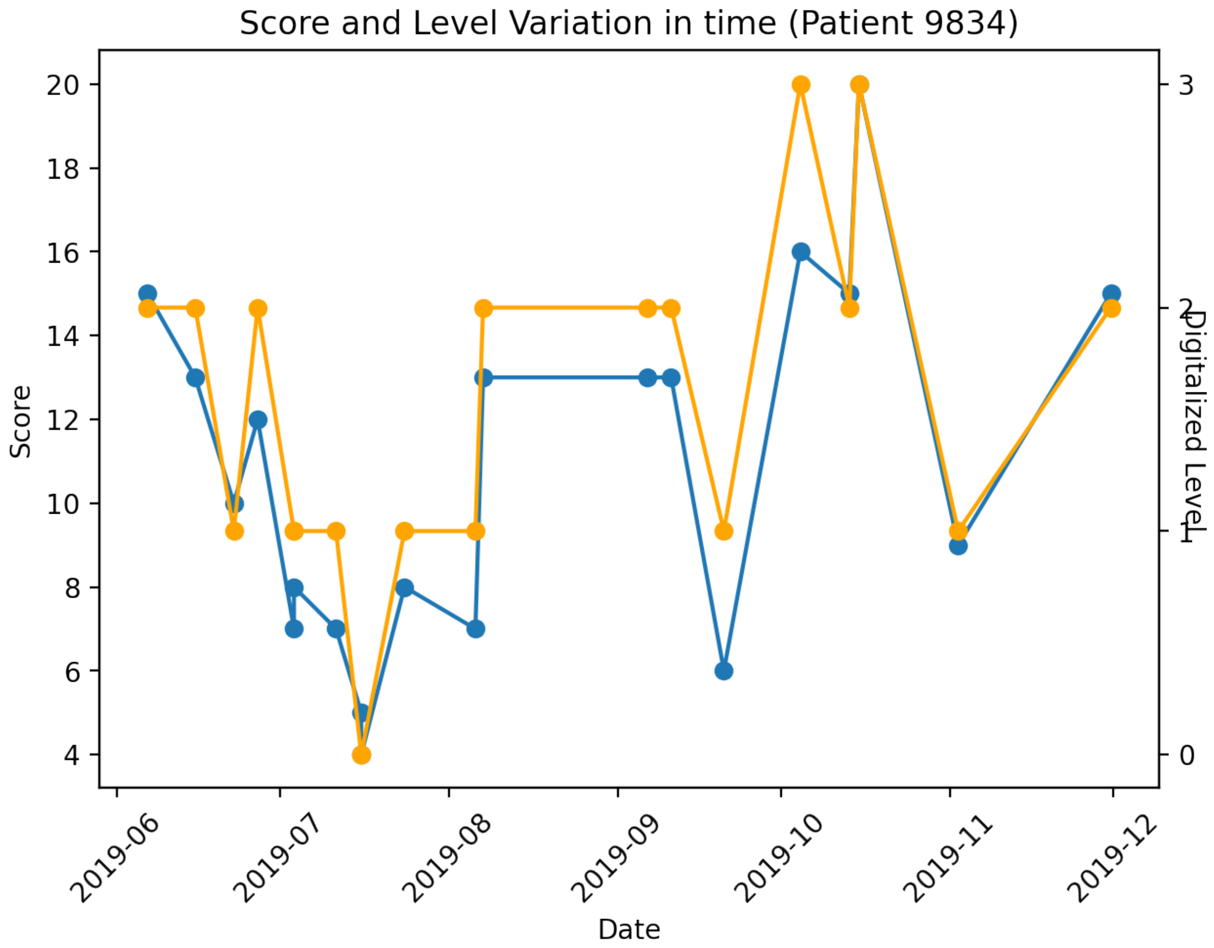
Visualization for progress of a certain patient

For visualization on a certain patient, I think the most straightforward way of doing so is by plotting line charts. This can clearly show to the patient his/her temporal progress. **Trend lines are the primary representation of the progress**. The following is just a preliminary visualization of a line chart.

But in fact, for real implementation, **an interactive visualzation** can be a lot more informative, we can only keep the variation in levels as the body, and use a **tooltip** to present information including the exact score, mediacal notes about following measures, analysis on current progress, and difference between the current result and the previous one. The information can then be checked and compared later by patients or providers easily with interactive operations.

In [10]:

```python
# Sample visualization of a patient's progress
pid = list(patient_logs.keys())[0]
sample = patient_logs[pid]
fig, ax1 = plt.subplots(figsize=(7,5), dpi=200)
ax2 = ax1.twinx()
ax1.scatter(sample['date'],sample['score'])
ax1.plot(sample['date'],sample['score'])
ax2.scatter(sample['date'],sample['level'],c='orange')
ax2.plot(sample['date'],sample['level'],c='orange')
plt.title("Score and Level Variation in time (Patient {})".format(pid)
)
plt.sca(ax1)
plt.xticks(rotation=45)
plt.xlabel("Date")
plt.ylabel("Score")
plt.sca(ax2)
plt.yticks((0,1,2,3))
plt.ylabel("Digitalized Level",rotation=270)
plt.show()
```



Other important information

Based on my analysis, I find several sets of information that are important to collect to help draw conclusions:

- Dates of treatment: In the current database, I can only see the date when a patient takes an assessment, and I have to assume the date of assessment is the day of treatment to draw the conclusions. If I can have the accurate days of treatment, then it'll be easier to analyze how the intervals between treatments really affect the development of a patient's condition along the timeline.
- Initial status of a patient: It's not clarified in this dataset, so I just assume that everyone is in his/her initial status when taking the first assessment. But the concept should be clearer so that I can analyze the effect of treatments on a patient from the earliest stage, which is more complete to reflect the development of a patient's mental condition.
- The information of treatment: In order to find out some features that can really help patients get better in the treatment, the information of treatment should also be collected so that regression analysis can be implemented to find out efficicient features that can help people recover from anxiety disorders.

As time is limited this time, I cannot exhaust all features I want to analyze, but if I can have more time, I'll first go to observe how the **frequency of assessments/treatments in a month** can affect the **variation in the assessment results**. It has the potential to get us a most suggested frequecy of taking an assessment.