U.S. Department of Transportation
Volpe Center

Training Provider Audit Prioritization Model
Data Analysis Findings and Proposal

Background

Risk is inherent in driving. On the road, drivers make decisions and unsafe decisions increase the risk of accidents. Drivers' training is intended to reduce the risk associated with unsafe driving decisions. By providing drivers the opportunity to absorb knowledge about factors that affect road safety (such as the vehicle, the road, and regulations) and to apply that knowledge through spending time "behind the wheel", drivers increase their ability to make safe decisions, which in turn reduces the likelihood of accidents on the road.

Today there are over 16,000 institutions registered as training providers responsible for training drivers who seek commercial driver licenses (CDL) in the U.S. Out of these, roughly half are also carriers whose revenue is tied to the shipments they perform in a particular time period. Both trainers and entry-level drivers have incentives to forego sufficient training prior to testing for the CDL and driving on the road: to bypass the time investment needed to train. While this may be to their immediate benefit, placing inadequately trained drivers on the road potentially increases the overall risk posed to the public.

An audit reduces the incentives to omit sufficient training by introducing the potential cost of negative audit results. Because of this, effective audits against trainers who do not provide adequate training reduce the number of inadequately trained drivers on the road, increasing its overall safety.

An audit prioritization model which assesses and ranks each training provider's risk levels help auditors optimize their resources, specifically by reducing the amount of time needed to process data and increasing the likelihood that audits are performed against providers who pose the greatest risk. The value of such model lies in its "predictiveness", or the degree to which it is able to describe each provider's risk. This ability hinges on the methodology and, ultimately, on the data used to quantify the risk associated with each provider.

This paper explores methodologies to build such a prioritization model using currently available data, namely data derived from FMCSA's Training Provider Registry (TPR). The purpose of the analysis is to demonstrate ways which best uses these available sources to model risk associated with training providers, thereby helping auditors prioritize their resources and improve road safety.

Risk

Risk is typically measured in terms of probability and severity. Risk levels of a particular event, for example, is estimated by the probability of its occurrence and severity of the results when it does occur. Entities, similar to events, are also typical subjects of risk assessments. For example, an insurance company may deem an entity as "high risk" if the entity frequently experienced severe losses in the past.

A training provider may be deemed as "higher risk" if, for example, data indicates that it trains more drivers annually or that it spends fewer hours training each driver compared to other providers. In this case, the quantity of drivers and the degree they under-train their drivers can be thought of as indicators of its risk severity. In the following sections we explore available data and assess ways to use its components to quantify risk associated with each training provider, with the end goal of ranking providers according to this basis.

Measure 1 - Training Hours

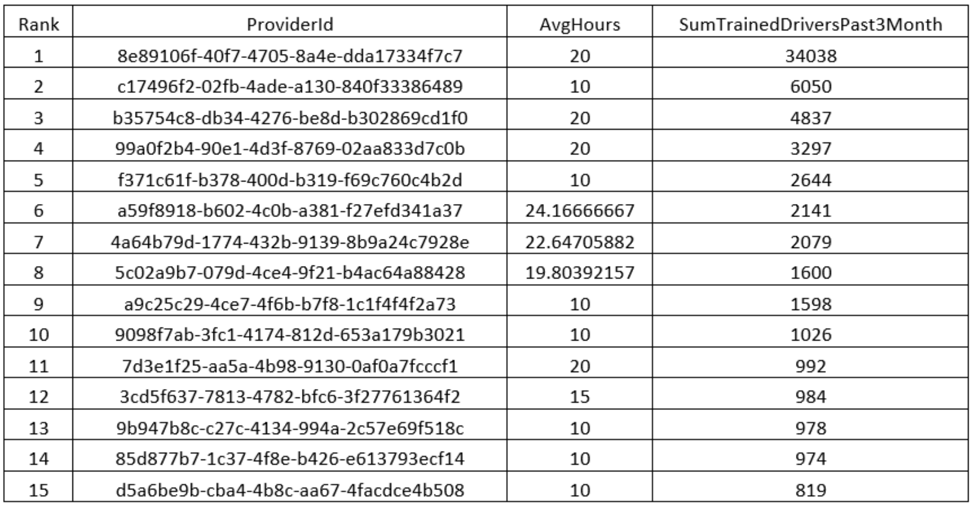
The TPR database stores data on time spent training drivers as inputted by the training providers. The database delineates between theory and "behind the wheel" hours, and providers input estimates of each at the time they register and as they certify each driver. Estimates of training hours per driver can be used as a measure for severity: trainers who perform fewer hours of training per driver can be deemed as under-equipping drivers with the knowledge and experience needed to make safe decisions. Auditors may combine this number by the number of drivers certified within a period to arrive at a more comprehensive measure of severity, and prioritize providers accordingly. For example, a provider which certify more drivers annually with below average training hours per driver may be ranked higher than one which certifies fewer drivers with average training hours, the former being deemed as posing the greater risk.

The training hours data is provider-inputted and can be falsified. Providers may misrepresent the number of hours spent on training, either as estimates provided during their registration process or as they certify each driver. This does not render it entirely purposeless as indicators of risk. Training providers who input low training hours can be assumed to have entered accurate data (providers who falsify data input high, not low, hours), therefore the subset of hours data which indicate low hours can be used to evaluate risk among those who inputted low training hours. The limitation in using training hours data in this manner is obvious: those who perform low hours of training will not be prioritized if they simply enter higher numbers in the registry.

Based on this approach, auditors may identify those who are providing low hours of training and rank them accordingly, either by using the hours data alone or combining it with other measures of severity, such as number of drivers certified annually. It is key to note that this is not a comprehensive ranking of the entire population of training providers, but a ranking among the subset of providers who input low hours of training (the universal estimate of risk based on this data cannot be reliably modeled as there may be higher risk providers that falsified their data and thus not reflected on this list).

Analysis of Current Data

Given a dataset where information on training hours exist, we may calculate for average hours spent training per driver and select a threshold which enables categorization of the training's adequacy. For example, we may define "low" as fewer than 30 hours of combined theory and behind-the-wheel training (with more data we may provide support to this assertion). We may then aggregate all providers who fall under this threshold, and rank them by the number of drivers they train within, for example, the past 3 months.



Rank	ProviderId	AvgHours	SumTrainedDriversPast3Month
1	8e89106f-40f7-4705-8a4e-dda17334f7c7	20	34038
2	c17496f2-02fb-4ade-a130-840f33386489	10	6050
3	b35754c8-db34-4276-be8d-b302869cd1f0	20	4837
4	99a0f2b4-90e1-4d3f-8769-02aa833d7c0b	20	3297
5	f371c61f-b378-400d-b319-f69c760c4b2d	10	2644
6	a59f8918-b602-4c0b-a381-f27efd341a37	24.16666667	2141
7	4a64b79d-1774-432b-9139-8b9a24c7928e	22.64705882	2079
8	5c02a9b7-079d-4ce4-9f21-b4ac64a88428	19.80392157	1600
9	a9c25c29-4ce7-4f6b-b7f8-1c1f4f4f2a73	10	1598
10	9098f7ab-3fc1-4174-812d-653a179b3021	10	1026
11	7d3e1f25-aa5a-4b98-9130-0af0a7fcccf1	20	992
12	3cd5f637-7813-4782-bfc6-3f27761364f2	15	984
13	9b947b8c-c27c-4134-994a-2c57e69f518c	10	978
14	85d877b7-1c37-4f8e-b426-e613793ecf14	10	974
15	d5a6be9b-cba4-4b8c-aa67-4facdce4b508	10	819

Figure 1: In this figure, providers who input fewer than 30 hours of training are ranked according to the number of drivers they trained in the last three months. This example uses data gathered from TPR registration and treats those who input below 30 hours as providing roughly equal amount of training to their drivers.

Measure 2 - Hours Over Cost

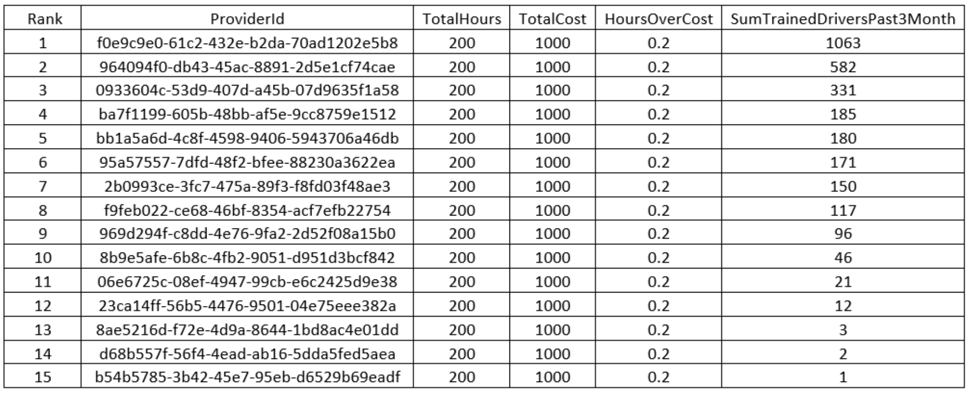
If a provider may falsify the amount of training they provide, it can be surmised that some who input very high training hours relative to other providers may do so fraudulently. However, high training hours alone may not necessarily indicate fraud as it is possible that a drivers perform high numbers of hours training in actuality. As a means to distinguish between the true and falsified inputs, we may compare the number of hours to the amount provider charges their drivers, and assume that trainers who input correspondingly high charges is more likely to actually provide the high training hours that indicated.

Auditors, in turn, may look at the ratio of training hours to charged cost, and rank those with highest hours-to-cost ratio as likeliest to inadequately train their drivers. Two assumptions are needed to rank providers in this manner. The first is that trainers who input high training hours and low tuition charges are likelier to be misrepresenting their numbers, and the second is that providers who misrepresent their numbers are likelier to provide inadequate training to their drivers. A more robust model can be built with data which enables testing of these assumptions, such as audit data verifying strong association between high hours-to-cost index and fraud. However, both assumptions are reasonable despite the absence of empirical support.

The hours-to-cost index is a measure of probability, not severity. How severely a provider is under-training their drivers cannot be estimated based on the components of the index, which are falsifiable. It is also important to note that the index is solely a *measure* of probability, not an estimate of the true probability. We use the index to estimate where a provider ranks in likelihood *in comparison to* other providers, while the true probability itself can only be estimated given more data, such as the verifying audit data mentioned previously.

Analysis of Current Data

An extension of the previous analysis, we may divide training hours by the cost charged by a provider to arrive at hours to cost ratio and aggregate those providers with the highest ratios. We may then rank providers based on the number of drivers they train within, for example, the past 3 months.



Rank	ProviderId	TotalHours	TotalCost	HoursOverCost	SumTrainedDriversPast3Month
1	f0e9c9e0-61c2-432e-b2da-70ad1202e5b8	200	1000	0.2	1063
2	964094f0-db43-45ac-8891-2d5e1cf74cae	200	1000	0.2	582
3	0933604c-53d9-407d-a45b-07d9635f1a58	200	1000	0.2	331
4	ba7f1199-605b-48bb-af5e-9cc8759e1512	200	1000	0.2	185
5	bb1a5a6d-4c8f-4598-9406-5943706a46db	200	1000	0.2	180
6	95a57557-7dfd-48f2-bfee-88230a3622ea	200	1000	0.2	171
7	2b0993ce-3fc7-475a-89f3-f8fd03f48ae3	200	1000	0.2	150
8	f9feb022-ce68-46bf-8354-acf7efb22754	200	1000	0.2	117
9	969d294f-c8dd-4e76-9fa2-2d52f08a15b0	200	1000	0.2	96
10	8b9e5afe-6b8c-4fb2-9051-d951d3bcf842	200	1000	0.2	46
11	06e6725c-08ef-4947-99cb-e6c2425d9e38	200	1000	0.2	21
12	23ca14ff-56b5-4476-9501-04e75eee382a	200	1000	0.2	12
13	8ae5216d-f72e-4d9a-8644-1bd8ac4e01dd	200	1000	0.2	3
14	d68b557f-56f4-4ead-ab16-5dda5fed5aea	200	1000	0.2	2
15	b54b5785-3b42-45e7-95eb-d6529b69eadf	200	1000	0.2	1

Figure 2: In this example, providers who claim high training hours but charge low cost are assumed to have higher likelihood of providing inadequate training, therefore they pose higher risk. With data that provides a benchmark for each provider's safety performance, this assumption can be tested and verified.

Measure 3 - Variability of Test Scores

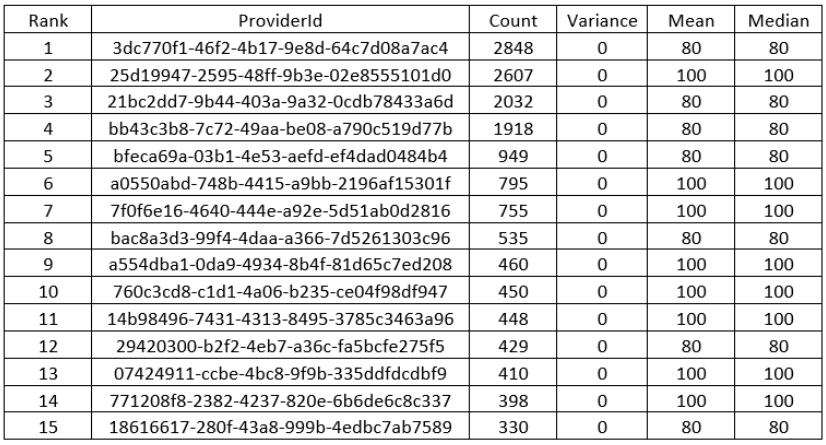
The TPR database stores drivers' theory test scores which are also recorded by providers and can also be falsified. The variance of such scores can be used as an indicator of whether the inputs are fabricated, and by extension, whether the provider is likely to provide inadequate training. For example, given 10 test-takers, a provider may input the unlikely outcome of a same score for each driver and consequently rank higher in the prioritization list. The assumption made in modeling risk in this manner is that providers who misrepresent their test scores have a higher likelihood to inadequately train, or to forgo actual training and testing altogether.

The probability that a certain set of test scores have been fabricated can be quantified given a distribution of test scores. For example, given 10 students taking a certain test, the probability that all scored a grade of 80 can be estimated given a known distribution of test scores. An auditor then may associate this unlikely test outcome as an indicator of higher likelihood of inadequate training, and rank providers with low-probability test outcomes as higher risk. The assumption needed in modeling risk in this manner is that providers who input unlikely outcomes (such as 10 drivers scoring the same score) are likelier to misrepresent their data, and those who do so are likelier to forgo the training, testing, or re-testing needed prior to driver certification.

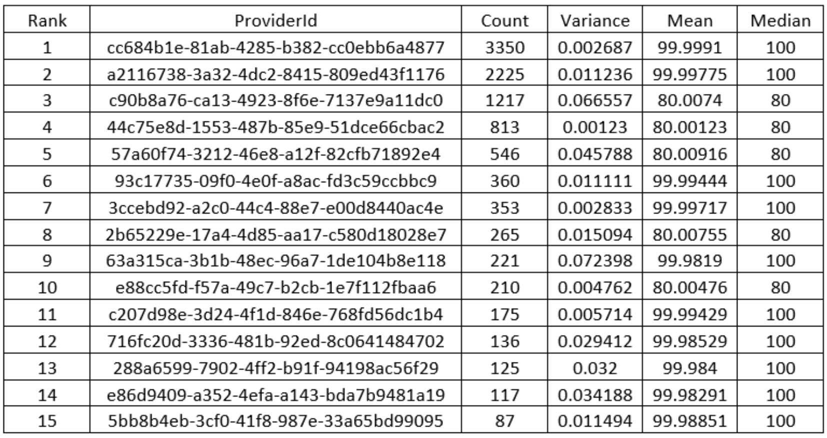
Analysis of Current Data

Given a known distribution of test scores and assuming independence among them, the probability of a certain outcome of test scores can be calculated from a given a list of scores. For example, given 3 test takers and each has 3 equally likely outcomes of A, B, and C, the probability that all score the same (zero variance) is 3/27. Precisely, there are 27 seven possible outcomes of test scores for the three students (AAA, AAB, AAC, ABA, ..., CCC) and out of these, 3 meet the criteria of having one distinct score (AAA, BBB, CCC).

We can likewise estimate the probability of a set of outcomes recorded in the TPR database, and use this as a likelihood indicator of under-training. The tables below list providers who inputs theory test scores with zero or slightly above zero variance, sorted by the count of drivers they have trained to date. Given the scores distribution gleaned from the database, the probability of such zero variance outcomes are effectively zero, making it almost certain that theory test scores have been falsified.



Rank	ProviderId	Count	Variance	Mean	Median
1	3dc770f1-46f2-4b17-9e8d-64c7d08a7ac4	2848	0	80	80
2	25d19947-2595-48ff-9b3e-02e8555101d0	2607	0	100	100
3	21bc2dd7-9b44-403a-9a32-0cdb78433a6d	2032	0	80	80
4	bb43c3b8-7c72-49aa-be08-a790c519d77b	1918	0	80	80
5	bfeca69a-03b1-4e53-aefd-ef4dad0484b4	949	0	80	80
6	a0550abd-748b-4415-a9bb-2196af15301f	795	0	100	100
7	7f0f6e16-4640-444e-a92e-5d51ab0d2816	755	0	100	100
8	bac8a3d3-99f4-4daa-a366-7d5261303c96	535	0	80	80
9	a554dba1-0da9-4934-8b4f-81d65c7ed208	460	0	100	100
10	760c3cd8-c1d1-4a06-b235-ce04f98df947	450	0	100	100
11	14b98496-7431-4313-8495-3785c3463a96	448	0	100	100
12	29420300-b2f2-4eb7-a36c-fa5bcfe275f5	429	0	80	80
13	07424911-ccbe-4bc8-9f9b-335ddfdcdbf9	410	0	100	100
14	771208f8-2382-4237-820e-6b6de6c8c337	398	0	100	100
15	18616617-280f-43a8-999b-4edbc7ab7589	330	0	80	80



Rank	ProviderId	Count	Variance	Mean	Median
1	cc684b1e-81ab-4285-b382-cc0ebb6a4877	3350	0.002687	99.9991	100
2	a2116738-3a32-4dc2-8415-809ed43f1176	2225	0.011236	99.99775	100
3	c90b8a76-ca13-4923-8f6e-7137e9a11dc0	1217	0.066557	80.0074	80
4	44c75e8d-1553-487b-85e9-51dce66cbac2	813	0.00123	80.00123	80
5	57a60f74-3212-46e8-a12f-82cfb71892e4	546	0.045788	80.00916	80
6	93c17735-09f0-4e0f-a8ac-fd3c59ccbbc9	360	0.011111	99.99444	100
7	3ccebd92-a2c0-44c4-88e7-e00d8440ac4e	353	0.002833	99.99717	100
8	2b65229e-17a4-4d85-aa17-c580d18028e7	265	0.015094	80.00755	80
9	63a315ca-3b1b-48ec-96a7-1de104b8e118	221	0.072398	99.9819	100
10	e88cc5fd-f57a-49c7-b2cb-1e7f112fbaa6	210	0.004762	80.00476	80
11	c207d98e-3d24-4f1d-846e-768fd56dc1b4	175	0.005714	99.99429	100
12	716fc20d-3336-481b-92ed-8c0641484702	136	0.029412	99.98529	100
13	288a6599-7902-4ff2-b91f-94198ac56f29	125	0.032	99.984	100
14	e86d9409-a352-4efa-a143-bda7b9481a19	117	0.034188	99.98291	100
15	5bb8b4eb-3cf0-41f8-987e-33a65bd99095	87	0.011494	99.98851	100

Figure 3: Zero or close to zero variances are highly improbable given distribution of test scores in TPR database

Conclusion

The current data is limited in its use to provide a comprehensive risk model of training providers in the United States. The training hours data provides some ability for auditors to prioritize providers based on the degree they under train their drivers, relative to other providers. The hours to cost The test scores data can be used to estimate the probability of a training provider is misrepresenting its scores, which can be interpreted as a likelihood indicator of inadequate training.

The inability to build a more rigorous risk assessment model from current data is primarily due to the lack of data verification and a "target variable" which can be used as an objective benchmark to assess each provider's risk levels. This variable, for example, can be a provider's CDL first-time pass rate, defined as the number of test takers who pass the CDL test on first attempt. In modeling risk, auditors may assess how factors such as training hours and test score results impact first-time pass rate target variable. Statistical analysis, for instance, may indicate that low test scores, and not low training hours, is more strongly associated with low pass rates. Such statistical model based on accurate data would allow auditors to take a more holistic assessment of a provider's risk-level (indicated by its pass rate), in this case allowing them an empirical basis to weigh test scores more heavily than training hours in ranking providers to be audited.

A perhaps more indicative target variable is a measure of the safety performance of the drivers themselves, which would allow direct measurement of how each previously discussed factors impact safety on the road. For example, given a database of first time CDL holders' performance within 3 years of obtaining license, an auditor may assess which providers or which factors associated with those providers are more strongly associated with the number of accidents in the first three years (a target variable which measures safety performance). Thus far we *assume* that low training hours and fraudulent scores are associated with higher safety risk, and with such information we may test if, for example, lower training hours are associated with higher average number of accidents within first 3 years of obtaining CDL. More generally, a target variable which measures safety performance allows analysis of how strongly each factor is related to overall road safety, which in turn allows a more comprehensive risk assessment of each training provider.

Building such a model requires that data on provider factors and target variable are both available and accurate. The current data set possess no viable target variable to measure factors against, and therefore there is no basis to decide how much each factor should weigh in estimating a provider's overall risk.

Data Improvements

Data improvement efforts should be aimed at allowing analysis into how provider factors affect the risk levels a provider poses. To illustrate, an ideal data set would be one which includes a set of provider factors that are indicative of risk levels (such as training hours, test scores, or CDL pass-rates) and a set of possible target variables which measures safety performance (such as number of accidents within the first three years of obtaining CDL). One concrete example is a table where each row describes one CDL holder and includes factors such as their provider, the driver's training hours, test score, and number of CDL test attempts and target variables such as number of accidents.

Data improvement efforts such as gathering new data or verifying inputs such as training hours and test scores may be prohibitively expensive. As such, the first step in improvement should be an inventory of existing data sources and an assessment of their integrity and suitability as indicators of risk, after which an analysis of relationships between each factor against target variables can be performed. From this a model which considers various aspects of a provider and outputs an estimate of the risk they impose can be built, and its overall effectiveness can be evaluated.