Pooled Testing: Determining The Optimum Pool Size To Minimize The Total Number Of Tests

Yonah Wilamowsky, Stillman School of Business, Seton Hall University
Viswa Viswanathan, Stillman School of Business, Seton Hall University
Sheldon Epstein, Stillman School of Business, Seton Hall University

ABSTRACT

In light of the rapidly spreading COVID-19 virus, the FDA has suggested pooling of samples in order to reduce the cost of testing a large population. Under this approach, several samples are pooled, and the pooled samples are first tested. If the pool tests negative, then the lab would have successfully tested many samples while consuming only the resources needed for a single test. If the pooled sample tests positive, then each sample that comprised the pool is individually tested. In this context, an important question for people in the field is "Given a certain overall infection rate among the population, what is the optimum pool size so that we can minimize the overall number of tests for a given number of individual samples?" In this paper, we derive this number both empirically and analytically. We also address the related question "Given a certain pool size, what is the maximum infection rate for which we can still gain in terms of the number of tests?"

Keywords: COVID-19; Pooled Sampling; Statistics; Optimization

INTRODUCTION

COVID -19 has affected virtually everyone on Earth and has touched nearly every aspect of life. As of this writing, more than six months into the pandemic, it has sickened more than 13 million people worldwide, and killed over 570,000. The United States has had over 3.5 million confirmed cases and 139,000 deaths, about 4.6% (New York Times, July 5, 2020). However, the World Health Organization has estimated that actual cases may be as many as ten times the number of confirmed cases. If so, it would place the fatality rate closer to 0.5%. The CDC has determined that the best way to diminish the spread of the virus is through testing and contact tracing.

To increase test coverage while simultaneously reducing the cost, the government is recommending pooled testing. In the form of pooled testing that we consider, each sample to be tested is divided into two parts and one part is put away safely. The remaining part of a group of samples (say, s of them) are combined into a pool and the pool is tested as a whole. If the pool tests negative, then all the samples that comprise the pool are cleared as negative, thereby enabling a single test to cover s samples. Instead, if the pool tests positive (which will happen if even one of the samples in the pool is positive), then the second part of each individual sample that formed the pool must be tested separately.

We can call the above method "two-stage pooled testing" because each sample could potentially need to be tested twice. Using this method would enable a university or business, for example, to test every individual on a regular basis (New York Times, July 1, 2020; Wall Street Journal, June 30, 2020). It would also make feasible the testing of hundreds of thousands, or even millions of people as needed. The FDA has published guidelines giving the technical requirements for pooled testing and has posted template updates regarding validation to be used for pooled samples (FDA, June 16, 2020).

It is quite clear that pooled testing will be most efficient when the percentage of the population infected is low. The higher the percentage infected, the greater the probability that the pool will test positive, and thus the less benefit there is. In fact, at a certain point, pooling will increase the total number of tests necessary rather than decreasing it.

Pooled testing has been widely studied in the biological sciences (see for example Pritchard & Tebbs, 2011 and Pilcher, Westreich & Hudgens 2020). To the best of our knowledge, the problem of finding the optimal pool size as a function of the infection rate has not been studied thus far. The purpose of this paper is to find the optimal pool size for a given population infection rate. We assume that dividing the original sample into two parts does not affect the reliability of the test. We present empirical and analytical approaches for finding the optimal pool size and also derive a formula for finding the highest infection level for which a given pool size yields some benefits by way of reduced number of tests.

EMPIRICAL SOLUTION

We first show an empirical method for finding the optimum pool size for a given infection rate.

We will use the following notation:

N: population size
p: percentage of population infected
s: pool size.

With the above notation in place, if a pool tests negative, then only one test need be done for the pool. If a pool tests positive, then each sample that formed the pool will be tested separately and hence require *s+1* tests in total.

The probability that a pool tests negative is (in Excel notation) BINOMDIST(0,s,p,0).

The probability that a pool tests positive is 1- BINOMDIST(0,s,p,0).

Therefore, the expected number of tests (*ntests)* that will be needed for a single pool of size s is:

$$BINOMDIST(0, s, p, o) + (s + 1) * (1 - BINOMDIST(0, s, p, 0)) \quad \text{(Eq 1)}$$

The total number of tests needed for a population of size N is N/s times this number. Table 1 shows the results for p between .01 and .50 and s between 2 and 20 for N = 1000.

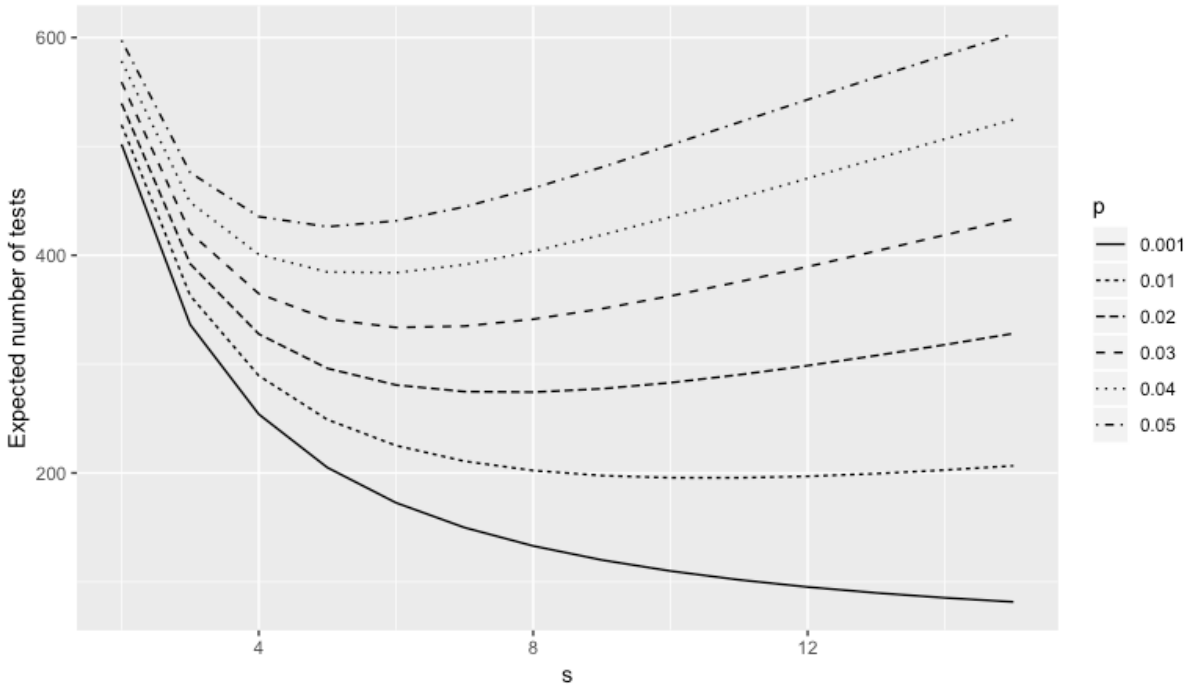
Table 1. Computing the pool size yielding the smallest number of overall tests for various infection rates (p) and pool sizes (s) for N = 1000

Expected Number of tests for each pool size/infection rate combination											
		Pool Size									
Min No of tests	**Opt pool size**	**p**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
196	11	0.01	520	363	289	249	225	211	202	198	196
274	8	0.02	540	392	328	296	281	275	274	277	283
334	6	0.03	559	421	365	341	334	335	341	351	363
384	6	0.04	578	449	401	385	384	391	404	419	435
426	5	0.05	598	476	435	426	432	445	462	481	501
466	5	0.06	616	503	469	466	477	494	515	538	561
502	4	0.07	635	529	502	504	520	541	565	591	616
534	4	0.08	654	555	534	541	560	585	612	639	666
564	4	0.09	672	580	564	576	599	626	655	683	711
594	4	0.10	690	604	594	610	635	665	853	724	751
719	3	0.15	778	719	728	756	790	822	957	879	903
821	3	0.20	860	821	840	872	905	933	1025	977	993
911	3	0.25	938	911	934	963	989	1009	1067	1036	1044
990	3	0.30	1010	990	1010	1032	1049	1061	1074	1071	1072
1005		0.31	1024	1005	1023	1044	1059	1068	1079	1076	1076
1019		0.32	1038	1019	1036	1055	1068	1076	1079	1080	1079
1033		0.33	1051	1033	1048	1065	1076	1082	1084	1084	1082
1046		0.34	1064	1046	1060	1075	1084	1088	1089	1087	1084
1050		0.35	1078	1059	1071	1084	1091	1094	1093	1090	1087
1050		0.40	1140	1117	1120	1122	1120	1115	1108	1101	1094
1050		0.45	1198	1167	1158	1150	1139	1128	1117	1107	1097
1050		0.50	1250	1208	1188	1169	1151	1135	1121	1109	1099

Expected Number of tests for each pool size/infection rate combination											
		Pool Size									
Min No of tests	**Opt pool size**	**p**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
196	11	0.01	196	197	199	203	207	211	216	221	226	232
274	8	0.02	290	299	308	318	328	339	350	360	371	382
334	6	0.03	376	389	404	419	433	448	463	478	492	506
384	6	0.04	453	471	489	507	525	542	559	576	592	608
426	5	0.05	522	543	564	584	603	622	641	658	675	692
466	5	0.06	585	607	630	651	671	691	710	727	744	760
502	4	0.07	641	665	688	709	730	749	768	785	801	816
534	4	0.08	691	716	739	760	780	799	817	833	848	861
564	4	0.09	737	761	783	804	824	841	858	872	886	898
594	4	0.10	777	801	823	843	861	877	892	905	918	928
719	3	0.15	924	941	956	969	979	988	996	1002	1007	1011
821	3	0.20	1005	1015	1022	1027	1031	1034	1036	1038	1038	1038
911	3	0.25	1049	1052	1053	1054	1053	1052	1051	1050	1048	1047
990	3	0.30	1071	1069	1067	1065	1062	1059	1056	1054	1051	1049
1005		0.31	1074	1072	1069	1066	1063	1060	1057	1054	1052	1049
1019		0.32	1077	1074	1070	1067	1064	1060	1057	1055	1052	1050
1033		0.33	1079	1075	1071	1068	1064	1061	1058	1055	1052	1050
1046		0.34	1081	1077	1072	1068	1065	1061	1058	1055	1052	1050
1050		0.35	1082	1078	1073	1069	1065	1061	1058	1055	1052	1050
1050		0.40	1087	1081	1076	1071	1066	1062	1059	1055	1053	1050
1050		0.45	1090	1083	1077	1071	1067	1062	1059	1056	1053	1050
1050		0.50	1090	1083	1077	1071	1067	1062	1059	1056	1053	1050

As expected, as p gets larger, the optimal pool size decreases. In fact, for any infection rate above 0.3 pooling will not reduce the total number of tests needed because the expected number of tests with pooling is in fact greater than testing singly; the probability of even a pool of size two being positive is 50% or more when p > 0.3. If the value of N is omitted (taken as 1), then the values produced would represent the expected number of tests as a proportion of the population size. Figure 1. shows the relationship from Table 1 graphically.

Figure 1. Relationship between pool size s and the expected number of tests for various values of p for a population of 1000



ANALYTICAL APPROACH

The total number of tests needed could be represented and solved analytically by the following approach.

Again, the expected number of tests for any pool (etpool) is the probability of the pool testing negative plus (s+1) times the probability of it testing positive. That is,

$$\begin{aligned} etpool &= (1-p)^s + (s+1)(1-(1-p)^s) \\ &= (1-p)^s + s(1-(1-p)^s) + (1-(1-p)^s) \\ &= 1 + s(1-(1-p)^s) \end{aligned} \quad \text{(Eq 2)}$$

This result can be thought of as follows: Every pool needs to be tested once whether positive or negative. This is represented by the first term of 1. For those groups that are positive, an additional s tests need to be done. This is given by the second term, which includes the probability that the pool tests positive.

The expected total number of tests for a population of size N (etpop) is:

$$\begin{aligned} etpop &= \left(\frac{N}{s}\right) etpool \\ &= \left(\frac{N}{s}\right)\left(1 + s(1-(1-p)^s)\right) \end{aligned} \quad \text{(Eq 3)}$$

Again, replacing the N with 1 gives the expected number of tests as a proportion of the population size.
In order to optimize the pool size s for a given p, we take the derivative and set to zero.

$$y = \left(\frac{1}{s}\right)\left(1 + s(1 - (1-p)^s)\right) \quad \text{(Eq 4)}$$

$$\frac{\partial y}{\partial s} = -log(1-p)(1-p)^s - \frac{1}{s^2} \quad \text{(Eq 5)}$$

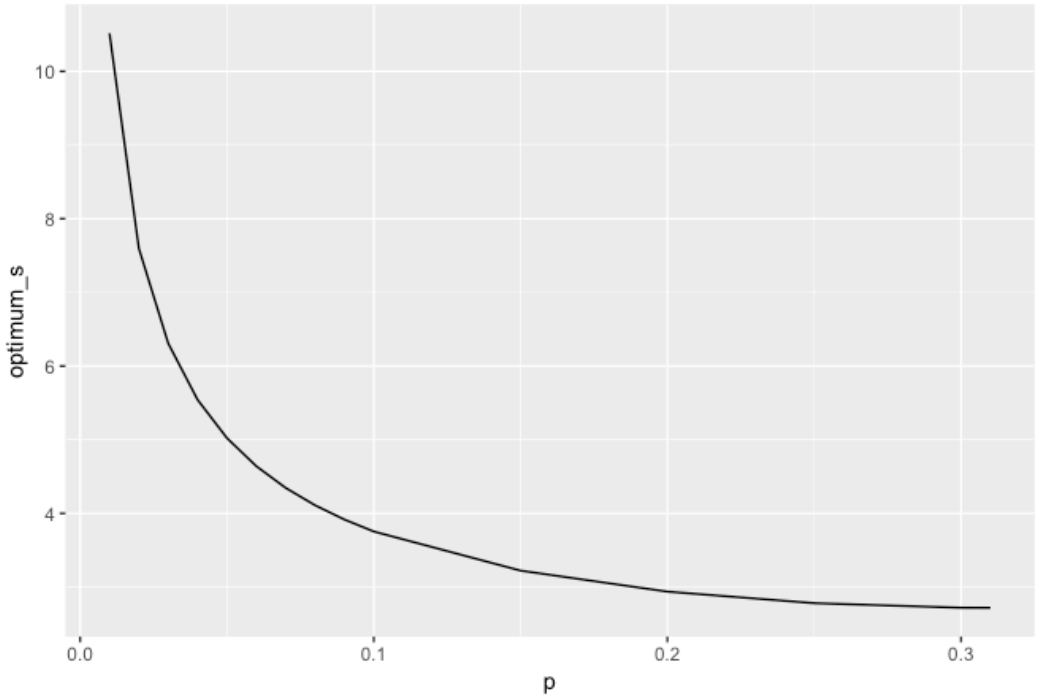
The root of this expression is:
$s = \frac{2W\left(\frac{i}{2}\sqrt{log(1-p)}\right)}{log(1-p)}$, where W is the Lambert Function. (Eq 6)

Table 2 shows the computed values of this function for some values of p, and Figure 2 shows the corresponding plot. If we round the optimum value of s, we can see that the results are consistent with the findings in Table 1. Beyond p = 0.31, the optimum value for s is 1 because any pooling beyond that increases the expected number of tests beyond N (as can be confirmed from Table 1).

Table 2. Optimum value of s for various values of p

p	optimum s
0.01	10.5162
0.02	7.59664
0.03	6.30753
0.04	5.54218
0.05	5.02239
0.06	4.64083
0.07	4.34619
0.08	4.11045
0.09	3.91682
0.1	3.75458
0.15	3.22329
0.2	2.93817
0.25	2.78175
0.3	2.71953
0.31	2.71838

Figure 2. Plot of the optimum value of s for various values of p

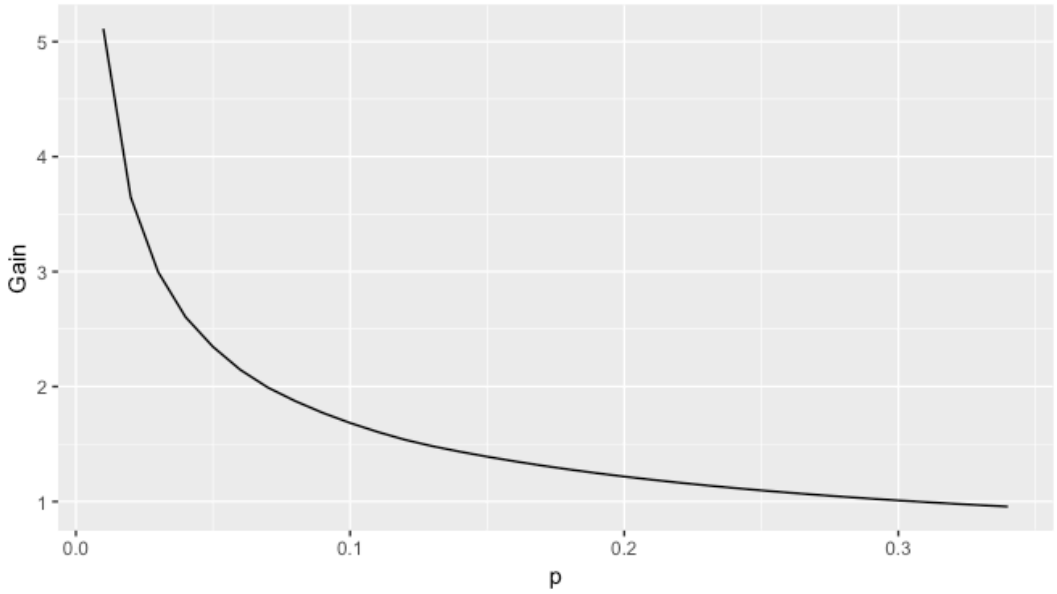


As mentioned before, the lower the proportion of positives in the population, greater are the benefits of pooling in the sense that we can test the entire population by using fewer tests than would be needed without pooling. If we define
$n_p = \ expected\ number\ of\ tests\ for\ an\ incidence\ rate\ of\ p\ using\ the\ optimum\ pool\ size$
Then, we can define the gain for a given incidence rate p as:

$$gain_p = \frac{N}{n_p} \quad \text{(Eq 7)}$$

Figure 3 plots gain against p. As expected, the gain goes below 1 at some point p greater than 0.3.

Figure 3. Gain as a function of p



In certain situations, it could be possible that there are external restrictions on the size of the pool. In this case, an important question might be "Given a value for s, what is the maximum value of p for which pooled testing provides at least some benefit in terms of reducing the number of tests?"
We can derive this as follows:

$$1 + s(1 - (1 - p)^s) < s$$
$$\Rightarrow p < 1 - \left(\frac{1}{s}\right)^{\left(\frac{1}{s}\right)} \quad \text{(Eq 8)}$$

The above is consistent with the numbers in Table 1. Table 3 shows the computed values of the maximum values for p for various values of s and highlights the fact that the largest number below 1000 in each column indeed corresponds to a probability that is less than the computed threshold.

Table 3. Maximum p (shown in the last row) for given values of s, such that pooling yields some benefit

Expected number of tests for each pool size/infection rate combination											
			Pool size								
Min no of tests	**Opt pool size**	**p**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
196	11	0.01	520	363	289	249	225	211	202	198	196
274	8	0.02	540	392	328	296	281	275	274	277	283
334	6	0.03	559	421	365	341	334	335	341	351	363
384	6	0.04	578	449	401	385	384	391	404	419	435
426	5	0.05	598	476	435	426	432	445	462	481	501
466	5	0.06	616	503	469	466	477	494	515	538	561
502	4	0.07	635	529	502	504	520	541	565	591	616
534	4	0.08	654	555	534	541	560	585	612	639	666
564	4	0.09	672	580	564	576	599	626	655	683	711
594	4	0.10	690	604	594	610	635	665	695	724	751
719	3	0.15	778	719	728	756	790	822	853	879	903
821	3	0.20	860	821	840	872	905	933	957	977	993
911	3	0.25	938	911	934	963	989	1009	1025	1036	1044
990	3	0.30	1010	990	1010	1032	1049	1061	1067	1071	1072
1005		0.31	1024	1005	1023	1044	1059	1068	1074	1076	1076
1019		0.32	1038	1019	1036	1055	1068	1076	1079	1080	1079
1033		0.33	1051	1033	1048	1065	1076	1082	1084	1084	1082
1046		0.34	1064	1046	1060	1075	1084	1088	1089	1087	1084
1050		0.35	1078	1059	1071	1084	1091	1094	1093	1090	1087
1050		0.40	1140	1117	1120	1122	1120	1115	1108	1101	1094
1050		0.45	1198	1167	1158	1150	1139	1128	1117	1107	1097
1050		0.50	1250	1208	1188	1169	1151	1135	1121	1109	1099
				Max p	0.293	0.307	0.293	0.275	0.258	0.243	0.229

Expected number of tests for each pool size/infection rate combination											
			Pool size								
Min no Of tests	**Opt pool size**	**p**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
196	11	0.01	196	197	199	203	207	211	216	221	226	232
274	8	0.02	290	299	308	318	328	339	350	360	371	382
334	6	0.03	376	389	404	419	433	448	463	478	492	506
384	6	0.04	453	471	489	507	525	542	559	576	592	608
426	5	0.05	522	543	564	584	603	622	641	658	675	692
466	5	0.06	585	607	630	651	671	691	710	727	744	760
502	4	0.07	641	665	688	709	730	749	768	785	801	816
534	4	0.08	691	716	739	760	780	799	817	833	848	861
564	4	0.09	737	761	783	804	824	841	858	872	886	898
594	4	0.10	777	801	823	843	861	877	892	905	918	928
719	3	0.15	924	941	956	969	979	988	996	1002	1007	1011
821	3	0.20	1005	1015	1022	1027	1031	1034	1036	1038	1038	1038
911	3	0.25	1049	1052	1053	1054	1353	1052	1051	1050	1048	1047
990	3	0.30	1071	1069	1067	1065	1062	1059	1056	1054	1051	1049
1005		0.31	1074	1072	1069	1066	1063	1060	1057	1054	1052	1049
1019		0.32	1077	1074	1070	1067	1064	1060	1057	1055	1052	1050
1033		0.33	1079	1075	1071	1068	1064	1061	1058	1055	1052	1050
1046		0.34	1081	1077	1072	1068	1065	1061	1058	1055	1052	1050
1050		0.35	1082	1078	1073	1069	1065	1061	1058	1055	1052	1050
1050		0.40	1087	1081	1076	1071	1066	1062	1059	1055	1053	1050
1050		0.45	1090	1083	1077	1071	1067	1062	1059	1056	1053	1050
1050		0.50	1091	1083	1077	1071	1067	1062	1059	1056	1053	1050
			0.217	0.206	0.196	0.187	0.179	0.172	0.165	0.159	0.154	0.148

CONCLUSIONS AND POLICY IMPLICATIONS

We have shown how to optimize pool size in two-stage pooled testing, the method that has been suggested by the FDA. Extensive testing and contact tracing are crucial in the effort to return to normalcy. As more and more segments of the economy start opening up, it becomes important to rapidly and cost-effectively test millions of samples. Our results have a very direct applicability in this context. As a suggestion for further study, we propose extending this approach to n-stage pooled testing in which when a pool tests positive on the first try, we do not automatically resort to testing each member of the pool separately; instead, we could do sub-pooling as well. Of course, this means that the quantity of material in each sample would need to be sufficiently large so as to be divided into many parts and still contain enough of the biological material to show up in the test if present. Another extension might consider additional costs of doing a pooled sample or time factors involved in waiting for multiple tests. Whatever efficiencies that can be introduced will hopefully help to end the scourge of the pandemic sooner.

AUTHOR BIOGRAPHIES

Dr. Yonah Wilamowsky is Professor of Computing and Decision Sciences at the Stillman School of Business of Seton Hall University. He teaches courses in statistics and operations research. His research and consulting interests center on applications of statistics and operations research to the law, medicine, business processes and higher education. His research has appeared in such journals as *Naval Research Logistics, Journal of the Operational Research Society, American Journal of Mathematical and Management Sciences, Property Tax Journal, Location Sciences* and *Computers and Operations Research.* E-mail: Yonah.wilamowsky@shu.edu (corresponding author)

Dr. Viswa Viswanathan is Associate Professor of Computing and Decision Science at the Stillman Schoool of Business of Seton Hall University in New Jersey. He teaches courses in Business Analytics and Information Technology. His research interests include Combinatorial Optimization, Online learning, Intelligent Tutoring Systems and Data Science. He has taught in Business Schools in India and the US and also worked in the software industry for a decade. His research has been published in Operations Research, IIE Transactions, IEEE Software and International Journal of AI in Education among others. He has also authored books on Data Science. E-mail: viswa.viswanathan@shu.edu

Dr. Sheldon Epstein is Professor of Computing and Decision Science of the Stillman School of Business of Seton Hall University. He teaches courses in Operations Management and Decision Making. His research has been published in a wide variety of Technical, Applied and Practitioner Journals including: *Computers & Operations Research, Journal of the Operational Research Society, The New York Statistician, Naval Research Logistics, Opsearch, American Journal of Mathematical and Management Sciences, Annals of the Society of Logistics Engineers, Journal of Property Tax Assessment and Administration, Property Tax Journal* and *Interface*. E-mail: Sheldon.epstein@shu.edu

REFERENCES

FDA (June 16, 2020). Coronavirus (Covid-19) Update: Facilitating Diagnostic Test Availability for Asymptomatic Testing and Sample Pooling, Retrieved on July 14, 2020 from https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-facilitating-diagnostic-test-availability-asymptomatic-testing-and

New York Times (2020, July 5). How Deadly is the Coronavirus: Scientists are Searching for a Definitive Answer, Retrieved on July 14, 2020 from https://www.nytimes.com/2020/07/05/world/coronavirus-updates.html#link-79d7c7d3.

New York Times (2020, July 1). Federal Officials Turn to a New Testing Strategy as Infections Surge, Retrieved on July 14, 2020 from https://www.nytimes.com/2020/07/01/health/coronavirus-pooled-testing.html

Pilcher, Christopher D, Westreich, Daniel, Hudgens, Michael G. "Group Testing for Sars-Cov-2 to Enable Rapid Scale-Up of Testing and Real-Time Surveillance of Incidence." The Journal of Infectious Diseases, jiaa378, https://doi.org/10.1093/infdis/jiaa378

Pritchard, Nicholas A, and Joshua M Tebbs. "Estimating Disease Prevalence Using Inverse Binomial Pooled Testing." Journal of agricultural, biological, and environmental statistics vol. 16,1 (2011): 70-87. doi:10.1007/s13253-010-0036-4

Wall Street Journal (June 30, 2020). Labs Turn to Pooled Testing for More Efficient Covid-19 Testing, Retrieved on July 14, 2020 from https://www.wsj.com/articles/labs-turn-to-pooled-testing-for-more-efficient-covid-19-surveillance-11593521544

NOTES