Against the F-score

Adam Yedidia

December 8, 2016

This essay explains why the F-score is a poor metric for the success of a statistical prediction.

1 What is the F-score?

From Wikipedia:

> In statistical analysis of binary classification, the F1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results, and r is the number of correct positive results divided by the number of positive results that should have been returned. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst at 0.
>
> The traditional F-measure or balanced F-score (F1 score) is the harmonic mean of precision and recall:
>
> $$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

2 Christmastime

Why use the F-score? Well, in my experience, the F-score comes under discussion as an alternative to just measuring the accuracy of some test that you're developing.

For example, imagine you're Santa Claus, and you have to distribute gifts to children around the world. Some children are nice, and others are naughty, and ideally, you'd like to give toys to the nice children and coal to the naughty children. Unfortunately, there are so many kids these days that you have trouble tracking them all. You know from past trends that about half of kids are nice, and about half are naughty. What do you do?

Well, you have a few options. One is to give every child toys. Another is to give every child coal. Alternately, you could flip a coin for each child, and randomly give each child toys or coal depending on the outcome of the flip. If you're interesting in maximizing the accuracy of your predictions, it doesn't matter what you choose to do—you'll end up giving half of children the right thing, and half of children the wrong thing.

This is where the F-score comes in. It is an alternative to simple accuracy, a possible thing-to-be-maximized. What would your behavior be if you decided to maximize F-score instead of accuracy?

Let's call giving a nice child toys a "true positive," and giving a naughty child coal a "true negative." Let's also call giving a naughty child toys a "false positive" and a nice child coal a "false negative." Now, remember what Wikipedia told us:

> [The precision] p is the number of correct positive results divided by the number of all positive results, and [the recall] r is the number of correct positive results divided by the number of positive results that should have been returned. (Note that we can define the accuracy in these terms as well, as the number of all correct results divided by the total number of results.)

Thus: the precision is:

$$\frac{\text{true positives}}{\text{false positives} + \text{true positives}}$$

and the recall is

$$\frac{\text{true positives}}{\text{false negatives} + \text{true positives}}$$

See Fig. 1 for a beautiful visual representation of the two values, also due to Wikipedia.

3 F-score vs. Accuracy

Now that we know how to compute the F-score in terms of true and false positives and negatives, let's use it to see what our F-score would be for each of our different strategies (give every child toys, give every child coal, or flip a coin).

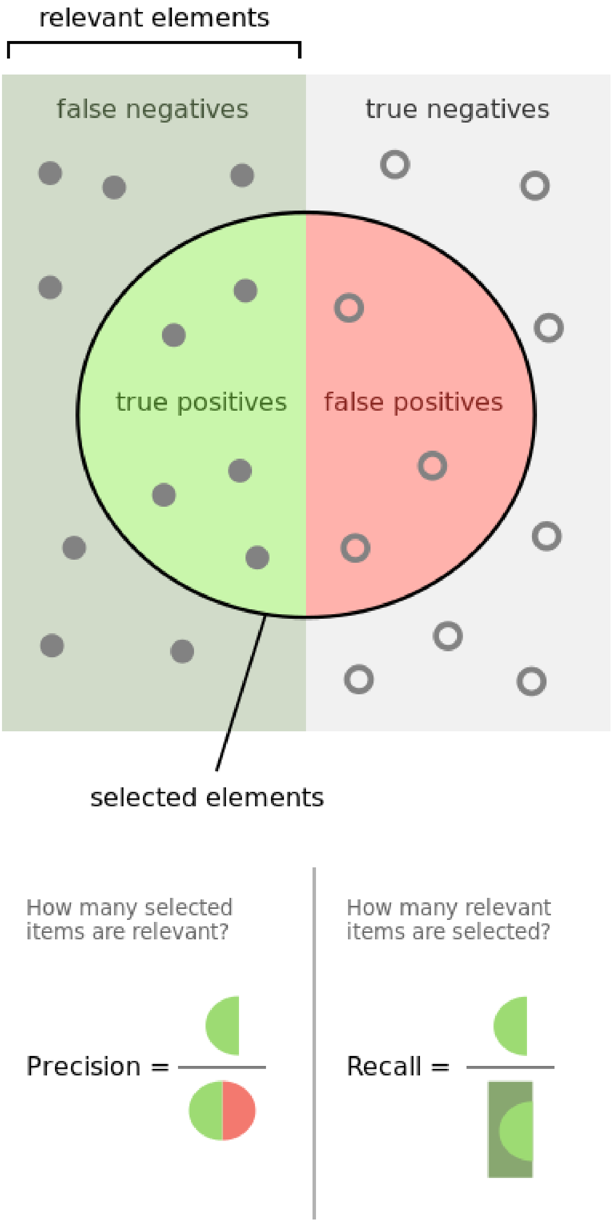

Figure 1: A visual representation of the meanings of the precision and recall. Figure due to Walber - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=36926283

Strategy	AP[1]	AN	TP	TN	FP	FN	Pre.	Rec.	F-score	Acc.
Toys for All	1	0	0.5	0	0.5	0	0.5	1	0.667	0.5
Coal for All	0	1	0	0.5	0	0.5	???	0	???	0.5
Flip a Coin	0.5	0.5	0.25	0.25	0.25	0.25	0.5	0.5	0.5	0.5

Now that we have the numbers in front of us, we can compare F-score to accuracy as a metric of success. Accuracy stubbornly told us that we did equally well no matter how many kids we gave toys to. What about the F-score? Well, by that metric, when we gave everyone toys, we did quite well. When we gave toys to only half of kids, we did a little less well. When we gave nobody toys, we... threw a divide-by-zero error as we tried to figure out how well we did. Not exactly a good start, eh?

But I'm a patient fellow. Let's see if we can't make this formula work in all cases with a little bit more tweaking. To that end, let's define the p-strategy to be the strategy where, for each child, we give them toys with probability p, and coal with probability $1-p$. We can easily rename our three strategies from before with this new framework: the "Toys for All" strategy is the 1-strategy, the "Coal for All" strategy is the 0-strategy, and the "Flip a Coin" strategy is the 0.5-strategy. By talking about our performance in terms of arbitrary p, we can add a new row to the table above:

Strategy	AP	AN	TP	TN	FP	FN	Pre.	Rec.	F-score	Acc.
p-strategy	p	$1-p$	$\frac{1}{2}p$	$\frac{1}{2}(1-p)$	$\frac{1}{2}p$	$\frac{1}{2}(1-p)$	$\frac{1}{2}$	p	$\frac{p}{\frac{1}{2}+p}$	$\frac{1}{2}$

Having rewritten the formula in terms of p (and done some cancellation in the process) it is now clear that the "Coal for All" strategy will yield an F-score of 0. Thus, according to the F-score, giving everyone toys is great, giving half of kids toys is still pretty good, but giving everyone coal is atrocious. The general behavior in terms of p is shown in Fig. 2.

This makes no sense. First, it seems strange that the "Coal for All" strategy should be given an F-score of 0, which is the worst possible. To my mind, it doesn't intuitively seem any worse than the other strategies discussed—and it is certainly not as bad a strategy as giving every child what they *don't* deserve, which gets the same F-score as "Coal for All".

[1] "AP" and "AN" refer to "All positives" and "All negatives," respectively. "TP" and "TN" refer to true positives and negatives, and "FP" and "FN" refer to false positives and negatives. "Pre." refers to the precision, "Rec." to the recall, and "Acc." to the accuracy.

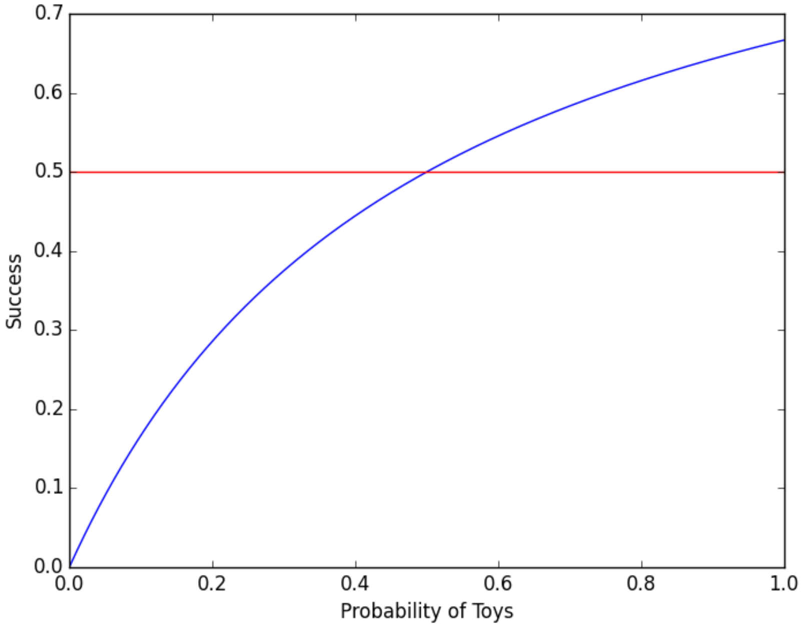

Figure 2: A plot of the success of a p-strategy in terms of p. "Success" is measured using accuracy in red, and using the F-score in blue.

The more serious problem, though, is that the behavior is asymmetric between positive results and negative results. In other words, I made the arbitrary decision earlier to call giving a child toys a "positive" result, and giving a child coal a "negative" result. Now, as a result of this entirely arbitrary choice, we are being told by the F-score that giving everyone toys is the best possible option, and that giving everyone coal is the worst thing since the Grinch stole Christmas! But if I'd called coal "positive" and toys "negative," the F-score would have immediately reversed its opinion.

Now, imagine if instead of half of kids being nice and half being naughty, if $\frac{1}{4}$ of them were nice and $\frac{3}{4}$ were naughty. Or, more generally, if q of them were nice and $(1-q)$ of them were naughty. Then, the table we know and love would look like this:

Strategy	AP	AN	TP	TN	FP	FN
p-strategy	p	$1-p$	pq	$(1-p)(1-q)$	$p(1-q)$	$(1-p)q$

Strategy	Pre.	Rec.	F-score	Acc.
p-strategy	q	p	$\frac{2pq}{p+q}$	$pq+(1-p)(1-q)$

Figs. 3 and 4 show the assessments of the accuracy, the F-score, and the "reversed" F-score—in other words, what the F-score would have given if I had reversed the labels for "positive" and "negative" examples.[2]Note that in both plots, the F-score gives a score of 0 to the strategy which maximizes accuracy—and steadily increases with decreasing accuracy. Between the fact that the F-score diverges so wildly from other intuitive metrics of success (such as the accuracy), and the fact that it changes so radically when you reverse the labels for "positive" and "negative" examples, it seems like the F-score is practically devoid of meaning.

4 Variable Scoring

At this point, the reader might object that I am using as my central example a situation in which the accuracy does fine on its own. The F-score is imperfect, one might argue, but in situations where the accuracy does even worse, it's worth using.

[2] Another way of describing the reversed F-score is as follows. If the F-score is the harmonic mean of the fraction of nice kids who got toys and the fraction of kids who got toys who were nice, the reversed F-score is the harmonic mean of the fraction of naughty kids who got coal and the fraction of kids who got coal who were naughty.

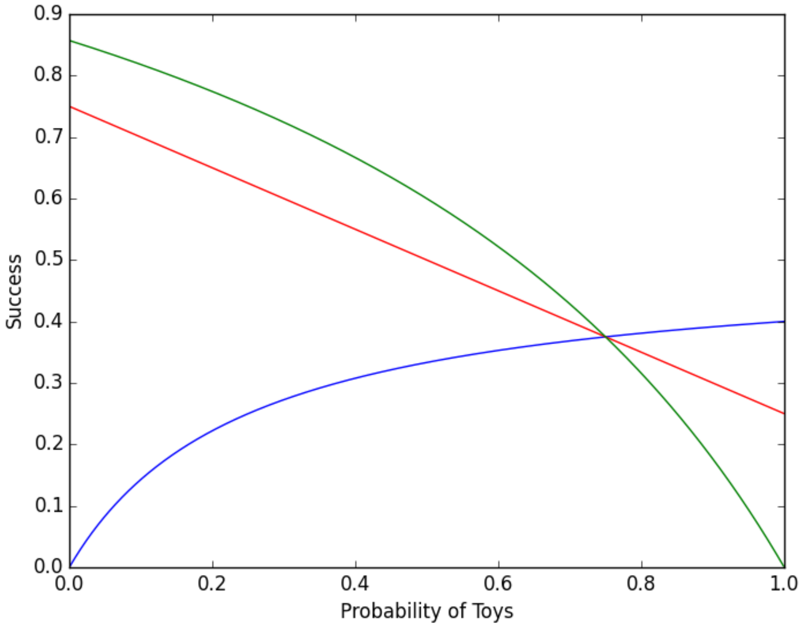

Figure 3: A plot of the success of a p-strategy in terms of p. “Success” is measured using accuracy in red, using the F-score in blue, and using the reversed F-score in green—in other words, what the F-score would have given if I’d reversed the “positive” and “negative” labels. In this plot, $q = 0.25$—in other words, $\frac{1}{4}$ of children are nice, and $\frac{3}{4}$ are naughty.

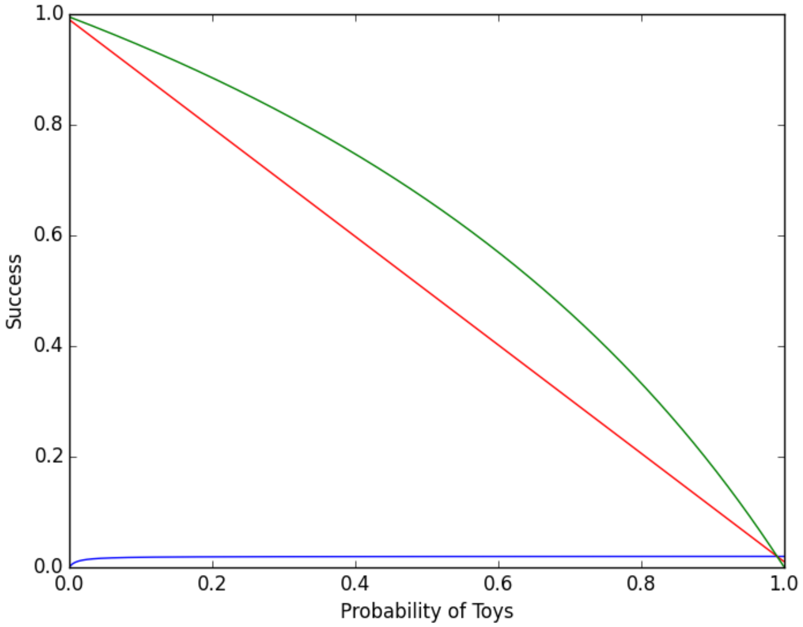

Figure 4: A plot of the success of a p-strategy in terms of p. "Success" is measured using accuracy in red, using the F-score in blue, and using the reversed F-score in green—in other words, what the F-score would have given if I'd reversed the "positive" and "negative" labels. In this plot, $q = 0.01$—in other words, $\frac{1}{100}$ of children are nice, and $\frac{99}{100}$ are naughty.

A prime example of a situation in which the accuracy fails as a good measure of success is a situation in which a false positive and a false negative are not equally bad. As an example, let us take the situation we were discussing before: $\frac{1}{4}$ of children are nice, and $\frac{3}{4}$ are naughty. The accuracy-maximizing strategy would of course be to give every child coal. But this isn't what Christmas is supposed to be about—Christmas is about good cheer! Giving every child coal doesn't achieve that.

So, in the spirit of Christmas, let us suppose that making sure that nice children receive toys is ten times more important than making sure that naughty children receive coal. If that's so, the accuracy-maximizing strategy is a miserable failure. It ensures that the 75% of children who are naughty get coal, but it misses the more important goal of ensuring that the 25% of children who are nice get toys.

How does the F-score-maximizing strategy do? Well, it depends on which one you're talking about. If giving a child toys is a "positive" event, then as we can see from Fig. 3, giving every child toys is the F-score-maximizing strategy—good! If, on the other hand, giving a child toys is a "negative" event, then giving every child coal is the F-score-maximizing strategy—too bad.

So yes, in this situation, the F-score may do better as a metric of success than the accuracy, depending on what exactly you're talking about. But really, neither of these metrics of success could possibly be expected to behave intelligently and responsively to the problem at hand, because we never told them about the fact that we cared more about the nice kids than about the naughty ones, or exactly how much more we cared. A fairer comparison would be between modified versions of the accuracy and F-score that accounted for this difference in importance somehow.

Enter the *weighted accuracy.* The weighted accuracy, instead of being:

$$\frac{\text{True Positives} + \text{True Negatives}}{\text{Total Examples}}$$

is now instead:

$$\frac{\alpha \cdot \text{True Positives} + \text{True Negatives}}{\alpha \cdot (\text{True Positives} + \text{False Negatives}) + \text{True Negatives} + \text{False Positives}}$$

for some parameter α which describes how much more important positive examples are than negative examples (in this case, how much more important nice children are than naughty ones).

But it would be unfair for me to adapt the accuracy formula for this difference in importance between children without doing the same for the F-score formula. How should the F-score formula be adapted? Once again, Wikipedia comes to the rescue:

> The general formula for positive real β is:
>
> $$F_\beta = (1+\beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}}$$
>
> The F-measure was derived so that F_β "measures the effectiveness of retrieval with respect to a user who attaches β times as much importance to recall as precision."

Now that we have two formulas which work for our new case, let's use them! In terms of p and q, the accuracy will go as:

$$\frac{\alpha pq + (1-p)(1-q)}{\alpha(pq + (1-p)q) + p(1-q) + (1-p)(1-q)}$$

which simplifies to:

$$\frac{\alpha pq + (1-p)(1-q)}{1 + q(\alpha - 1)}$$

The F-score, in contrast, will go as:

$$\frac{pq(1+\beta^2)}{\beta^2 q + p}$$

Finally, the "reversed" F-score (in which we call giving kids toys "negative" and coal "positive"), will go as:

$$\frac{(1-p)(1-q)(1+\frac{1}{\beta^2})}{\frac{1-q}{\beta^2} + 1 - p}$$

(Note the inversion of the β factor, to indicate that in this new regime, we consider positives to be $\frac{1}{\beta}$ times as important as negatives.)

We've decided that nice children are 10 times more important than naughty ones; that gives us $\alpha = \beta = 10$. Figs. 5 and 6 shows what these three metrics tell us to do for $q = 0.25$ and $q = 0.01$, respectively (the cases in which 25% and 1% of children are nice).

As we can see from Figs 5 and 6, the weighted accuracy does what one would expect: rewarding strategies that give children toys when $q = 0.25$

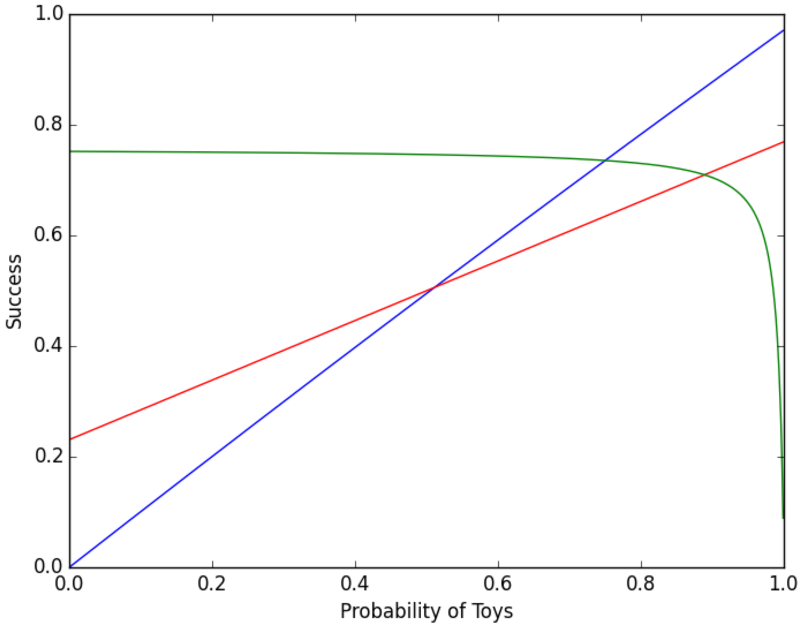

Figure 5: A plot of the success of a p-strategy in terms of p. “Success” is measured using accuracy in red, using the F-score in blue, and using the reversed F-score in green—in other words, what the F-score would have given if I’d reversed the “positive” and “negative” labels, as well as inverted the relative importance of positive instances to negative ones. In this plot, $q = 0.25$—in other words, $\frac{1}{4}$ of children are nice, and $\frac{3}{4}$ are naughty. Additionally, $\alpha = \beta = 10$, which means that we care about giving nice children what they deserve ten times more than we care about giving naughty children what they deserve.

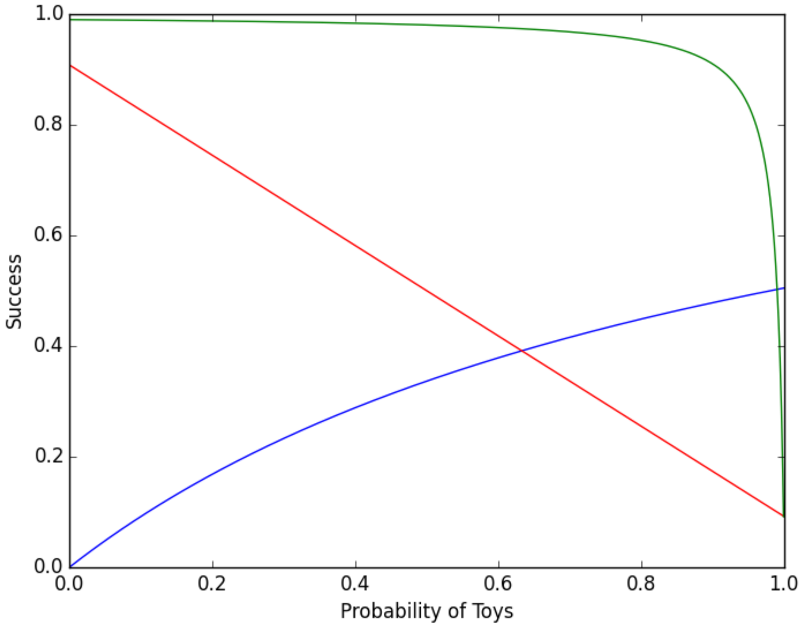

Figure 6: A plot of the success of a p-strategy in terms of p. "Success" is measured using accuracy in red, using the F-score in blue, and using the reversed F-score in green—in other words, what the F-score would have given if I'd reversed the "positive" and "negative" labels, as well as inverted the relative importance of positive instances to negative ones. In this plot, $q = 0.01$—in other words, $\frac{1}{100}$ of children are nice, and $\frac{99}{100}$ are naughty. Additionally, $\alpha = \beta = 10$, which means that we care about giving nice children what they deserve ten times more than we care about giving naughty children what they deserve.

and the bulk of the importance is contained in the nice kids, and rewarding strategies that give children coal when $q = 0.01$ and the bulk of the importance is contained in the naughty kids. The weighted F-score, in contrast, continues to display erratic and largely inscrutable behavior. Once again, the F-score and reversed F-score give opposite recommendations (which is alarming, given that they differ based only on an arbitrary choice of labelling). It seems like for the most part, which of these two you choose is more important than the choice of β or even the problem's parameters like q—which is terrible, because it means you have to know the answer you're looking for before you ask the question.

5 Conclusion

In conclusion, the F-score is a poor measure of success, on several levels. It seems like a clever idea at first glance, but the asymmetry between positive and negative examples buried within its formula makes its interpretation entirely dependent on an arbitrary choice of labels. This, to my mind, is its most damning flaw.

In addition, whereas the accuracy formula is simple and intuitive, with a short English-language explanation of what it does ("the fraction of children who got what they deserved"), the F-score formula has no correponding easy explanation. The closest I can come to is "the harmonic mean of the fraction of nice children who got toys, and the fraction of children who got toys who were nice," which is a lot harder to keep in your head at once, and requires you to have an intuition for what the "harmonic mean" is. (Is there anybody who knows the formula for the harmonic mean off the top of their head?).

Perhaps the strongest praise I can give to the F-score is that giving one more child what she deserves, instead of what she doesn't deserve, will never cause the F-score to decrease. (It won't necessarily cause it to increase, though.) This is much weaker than what could be said about the accuracy, however, and is an extremely weak endorsement in general. In particular, it is weak enough praise that it also applies to the Point Three Score, which is a metric of a test's success devised by me. Given a strategy for predicting positive and negative results, it gives that strategy a score of 0.3. Just like the F-score, the Point Three Score will never decrease in response to giving one more child what she deserves instead of what she doesn't deserve. Unlike the F-score, however, reversing the "positive" and "negative" labels on your examples won't change the Point Three Score's assessment of strategy. Finally, the Point Three Score is guaranteed to lie between 0 and 1 (like the accuracy and F-score), which is nice.

Therefore, dear reader, I recommend that if you are looking for an alternative to the accuracy or the weighted accuracy, that you consider using the Point Three Score before you use the F-score. I consider its assessments of a strategy's success to be more meaningful than those of the F-score.