A Note on Improving the Measurement of the Quality of Forecasts in Prediction Tournaments

Benedikt Alexander Schuler
benedikt.schuler@unisg.ch

Institute of Management and Strategy
University of St. Gallen
Switzerland

Johann Peter Murmann
Peter.Murmann@unisg.ch

Institute of Management and Strategy
University of St. Gallen
Switzerland

Marie Beisemann
beisemann@statistik.tu-dortmund.de

Department of Statistics
TU Dortmund University
Germany

9 November 2020

Abstract

We begin by arguing that the quality of forecasts consists of two aspects: accuracy and timing. Existing conceptualizations and operationalizations of the quality of forecasts, however, appear to focus more on the accuracy and do not incorporate sufficiently the timing of forecasts. To improve the so-called *Accuracy Score* of Cultivate Labs (the backend of GJopen and our St. Gallen Forecasting platform), we propose what we call the "*Quality Score",* which also considers the timing of bad forecasts. We believe that Cultivate Labs should consider adopting a *Quality Score* instead of its current *Accuracy Score*. Building on Merkle et al. (2017), we then move into a second important proposal by noting that research is frequently interested in measuring as precisely as possible the *forecasting skills* of persons. We argue that IRT models should be the preferred tool for measuring the forecasting skills of persons as they usually allow researchers to measure the forecasting skills more accurately than the *Accuracy Score* or *Quality Score*. To allow researchers to estimate forecasting skills, we refine an earlier IRT model to implement our definition of the quality of forecasts in the context of forecasting tournaments. Unlike earlier IRT models, which only captured the timing of one forecast per tournament question, our proposed model makes it possible to assess the timing of multiple forecast per tournament question, as is common on the GJopen and St. Gallen forecasting platforms. With our refined IRT model one can analyze experimental settings in which forecasters are encouraged to update their probability forecasts every time they obtain relevant new information.

Introduction

"Seeing the future sooner" than your competitors can give you a crucial competitive advantage in business. For this reason, scholars have begun to investigate how businesses can make high-quality forecasts to become more successful (Schoemaker & Tetlock, 2016). Although the quality of forecasts can be defined in multiple ways, a concise and intuitive definition of the quality of forecasts comprises two aspects: accuracy and timing. Predicting, for example, industry sales for the next year with an error of 1% is a better forecast than making the prediction with an error of 10%. Similarly, predicting sales with 1% error 365 days before the numbers are officially published is better than making a prediction with this accuracy 5 days before. The second forecast had 360 days of additional information available, making it much easier to give an accurate forecast. Existing conceptualizations and operationalizations of the quality of forecasts, however, appear to focus more on the accuracy and do not incorporate sufficiently the timing of forecasts. While Cultivate Labs (the backend company of GJopen and our St. Gallen forecasting platform) attempts to incorporate the timing of good forecasts into its measurement of forecast quality, we argue that its measurement can be improved upon because it does not adequately consider the timing of bad forecasts. The goal of this research note is threefold: First, we will provide a definition of what the quality of forecasts encompasses. Second, we review existing operationalizations of *forecasting quality* as well as related concepts and show that to date there is not a full implementation of the idea of *forecasting quality* in existing operationalizations because the timing of the forecasts is usually not captured as intended. This leads us to make a proposal on how to operationalize the quality of forecasts better. Third, we argue that *Item Response Theory* models are well suited to measuring the forecasting skills of a person and, therefore, refine an *Item Response Theory model* to implement to our definition of forecasting quality.

Defining the Quality of Forecasts

Although the quality of forecasts can be defined in multiple ways, the motto of the Good Judgment Project "See the Future Sooner" (https://goodjudgment.com/) provides a concise and intuitive definition of the quality of forecasts: a good forecast is an accurate prediction about a future state of the world ("seeing the future") that was made earlier than other good forecasts ("sooner"). This corresponds to the underlying rationale of the example in the introduction: Predicting industry sales for the next year with an error of 1% 365 days before the numbers are officially published is a better forecast than making a prediction with this accuracy 5 days before the numbers are officially published. The second forecast had 360 days with additional information available, making it much easier to give an accurate forecast. Likewise, the inaccuracy of a forecast has a temporal dimension. Drawing on the same example, it means that predicting industry sales for the next year with an error of 10% 5 days before the numbers are published is a worse forecast than making a prediction with this error 365 days before the numbers are officially published. The second forecast faced much more uncertainty because many things could intervene in the next 365 days whereas the first 5-day forecast would know how 360 out of the 365 days of the year had gone. Taken together, this means that the quality of forecasts consists of two aspects: accuracy and timing.

This reasoning is built upon the assumption that an event is easier to forecast when the forecaster is closer to the realization of the future state of the world on a temporal dimension, as at this point there should be more relevant information on the future state of the world available (this assumes the increasing availability of relevant information over time). Relating this once again to the example in the introduction, this means that more relevant information on the industry sales for the next year, such as more and better data on customer or market trends relevant to the

industry, should be available when the forecast is made 5 days before the numbers are officially published compared to when the forecast is made 365 days before the numbers are officially published. The assumption of the increasing availability of relevant information over time implies that the timing of a forecast (i.e., when a forecast was made) should be considered if the quality of a forecast is to be evaluated.

To better understand the meaning of accuracy and timing, let us consider the example depicted in Table 1. Persons 1 and 2 made more accurate forecasts than persons 3 and 4 as they assigned higher probabilities to the actual future state of the world at day 3 "Yes". Therefore, persons 1 and 2 should be ranked higher than persons 3 and 4.

Table 1: Example Ranking of Forecasters Based on the Accuracy and Timing of their Forecasts (ATF)

Persons	Answers	Day 1	Day 2	Day 3	Rank_{ATF}
Person 1	Yes	0.75	0.75	Yes	1
	No	0.25	0.25		
Person 2	Yes		0.75	Yes	2
	No		0.25		
Person 3	Yes	0.25	0.25	Yes	3
	No	0.75	0.75		
Person 4	Yes		0.25	Yes	4
	No		0.75		

Note. Answers = Possible future states of the world forecast by each person—a higher probability assigned to an answer means that a person deems this answer more likely than other answers; Day 1 = Probabilities that each person assigned to each possible future state of the world at day 1 (= probability forecasts at day 1); Day 2 = Probabilities that each person assigned to each possible future state of the world at day 2 (= probability forecasts at day 2); Day 3 = True future state of the world that was forecast; $\mathbf{Rank}_{\boldsymbol{ATF}}$ = Expected rank of each person based on their predictions at day 1 and day 2 if the quality of forecasts is conceptualized using the aspects accuracy and timing.

Furthermore, person 1 made their forecast one day earlier than person 2, and hence the forecast of person 1 was timelier than the forecast of person 2. Under the assumption that an event is easier to forecast when the forecast is closer to the realization of the future state of the world on

a temporal dimension as there should be more relevant information on the future state of the world available (assumption of the increasing availability of relevant information over time), the forecast at day 1 should have been more difficult than the forecast at day 2. Therefore, person 1 should be ranked higher than person 2. Similarly, even though both person 3 and person 4 did not assign high probabilities to the actual future state of the world at day 3 "Yes", person 3 made their forecast one day earlier than person 4 — that is, their forecast was timelier than the forecast of person 4. Under the assumption of the increasing availability of relevant information over time, the forecast at day 1 should have been more difficult than the forecast at day 2. Therefore, person 3 should be ranked higher than person 4.

Measuring the Quality of a Forecast

Having defined what the quality of forecasts encompasses, we are now going to describe different operationalizations of the quality of forecasts and related concepts used by Mellers and colleagues (Mellers et al., 2014, 2015, 2019) and by the Cultivate platform on which GJopen runs. These operationalizations are going to be evaluated against their strengths, their weaknesses, and their fit with the conceptualization of the quality of forecasts as outlined above. Based on this evaluation, we will introduce a new operationalization of the quality of forecasts that we call the *Quality Score*.

Mellers and colleagues. Mellers and colleagues (Mellers et al., 2014, 2015, 2019) equate the quality of forecasts with the accuracy of forecasts. Consequently, they operationalize forecasting quality by using *Brier Scores* which are formally defined as

$$BS_{iq} = \frac{1}{T_{iq}} \sum_{t=1}^{T_{iq}} \sum_{c=1}^{C} (f_{iqtc} - o_{qc})^2,$$

where BS_{iq} is the *Brier Score* of person i ($i \in \{1, \dots, N\}$) on question q ($q \in \{1, \dots, Q\}$); T_{iq} is the number of time-units (usually days) $t = 1, \dots, T_{iq}$ person i had an active forecast on question q; t indicates the number of time units (usually days) until the resolution of question q; C is the number of possible disjunct classes c ($c \in \{1, \dots, C\}$) the event can fall into; f_{iqtc} is the probability forecast of person i on question q at time t for class c; and o_{qc} indicates whether the class c is the true state of the world at the time of the resolution of question q ($o_{qc} = 1$ if yes, $o_{qc} = 0$ if not). The formal definition of the *Brier Score* reveals that it is a mean squared error—that is, the *Brier Score* measures the deviation of the forecasts f_{iqt} of person i on question q at time t from the true state of the world when question q is resolved. Hence, the *Brier Score* measures how accurately person i forecast the future state of the world on average. The *Brier Score* can assume values between 0 and 2, where values closer to 0 indicate better forecasts; values of 0.5 are neutral; and values closer to 2 indicate worse forecasts. The *Brier Score* is appropriate for binary and categorical outcomes that can be structured as true or false.

Although *Brier Scores* represent an adequate measure for the accuracy of a forecast, they do not take the timing of a forecast into account. Table 2 clearly illustrates this issue: Although person 1 made their forecast one day earlier than person 2, the *Brier Score* of person 1 equals the *Brier Score* of person 2. Similarly, even though person 3 made their forecast one day earlier than person 4, the *Brier Score* of person 3 equals the *Brier Score* of person 4. Under the assumption of the increasing availability of relevant information over time, person 1 should be ranked higher than person 2 and person 3 should be ranked higher than person 4. However, it is important to note that *Brier Scores* do represent an adequate operationalization of the accuracy of forecasts, which is the primary focus in the research by Mellers and colleagues (Mellers et al., 2014, 2015, 2019).

Table 2: Comparison of the Forecasters' Ranks based on the ATF and Brier Score

Persons	Answers	Day 1	Day 2	Day 3	Rank_{ATF}	Brier Score	Rank_{BS}
Person 1	Yes	0.75	0.75	Yes	1	0.125	1.5
	No	0.25	0.25				
	DBS	0.125	0.125				
Person 2	Yes		0.75	Yes	2	0.125	1.5
	No		0.25				
	DBS		0.125				
Person 3	Yes	0.25	0.25	Yes	3	1.125	3.5
	No	0.75	0.75				
	DBS	1.125	1.125				
Person 4	Yes		0.25	Yes	4	1.125	3.5
	No		0.75				
	DBS		1.125				

Note. Answers = Possible future states of the world forecast by each person—a higher probability assigned to an answer means that a person deems this answer more likely than other answers; DBS = Daily Brier Score—the Brier Score calculated at the specific day when a forecast was made; Day 1 = Probabilities that each person assigned to each possible future state of the world at day 1 (= probability forecasts at day 1); Day 2 = Probabilities that each person assigned to each possible future state of the world at day 2 (= probability forecasts at day 2); Day 3 = True future state of the world that was forecast; $\mathbf{Rank}_{\boldsymbol{ATF}}$ = Expected rank of each person based on their predictions at day 1 and day 2 if the quality of forecasts is conceptualized using the aspects accuracy and timing; $\mathbf{Rank}_{\boldsymbol{BS}}$ = Rank based on the Brier Score.

Although *Brier Scores* do not represent an adequate operationalization of the quality of forecasts as conceptualized above, *Brier Scores* can provide the basis for scores measuring both the accuracy and timing of a forecast. One example of such a score is the *Accuracy Score* used by the Cultivate forecasting platform.

The Accuracy Score by Cultivate. Cultivate operationalizes the quality of forecasts by using the *Accuracy Score*[1]. The *Accuracy Score* builds upon the *Daily Brier Score* of person i on question q at time t which is formally defined as

[1] Cultivate uses different names for the same score: The *Accuracy Score* is also called *Net Brier Score*.

$$DBS_{iqt} = \sum_{c=1}^{C} (f_{iqtc} - o_{qc})^2$$

and assigns each person i a *Brier Score* for each day (time unit) t they had an active forecast on question q.

The *Accuracy Score* can then formally be defined as

$$AS_{iq} = \left[\frac{1}{T_{iq}} \sum_{t=1}^{T_{iq}} DBS_{iqt} - \frac{1}{T_{iq}} \sum_{t=1}^{T_{iq}} median\left(DBS_{1qt}, \dots, DBS_{N_{qt}qt}\right) \right] \cdot p_{iq}$$

$$= \left[\frac{1}{T_{iq}} \sum_{t=1}^{T_{iq}} DBS_{iqt} - median\left(DBS_{1qt}, \dots, DBS_{N_{qt}qt}\right) \right] \cdot \frac{T_{iq}}{T_q}$$

$$= \frac{1}{T_q} \sum_{t=1}^{T_{iq}} DBS_{iqt} - median\left(DBS_{1qt}, \dots, DBS_{N_{qt}qt}\right),$$

where AS_{iq} is the *Accuracy Score* of person i on question q; $p_{iq} = \frac{T_{iq}}{T_q}$ is the participation rate indicating the percentage of days forecaster i had an active forecast on question q; T_q is the number of time units (usually days) $t = 1, \dots, T_q$ question q was forecastable; and N_{qt}, which is part of the subscript of the DBS, indicates the number of persons i $\left(i \in \{1, \dots, N_{qt}\}\right)$ who made a forecast on question q at time t. The *Accuracy Score* can assume values between -2 and 2, where negative values indicate that the forecasts of a person were, overall (i.e., across the days a person made a forecast), more accurate than the typical forecast at each day; a value of 0 indicates that the forecasts of a person were, overall, as accurate as the typical forecast at each day; and positive values indicate that the forecasts of a person were, overall, less typical than the average forecast at each day.

Although the *Accuracy Score* takes the timing of good forecasts into account,[2] it does not adequately consider the timing of bad forecasts, and thus distorts the measurement of the quality of forecasts as conceptualized above.[3]

Table 3: Comparison of the Forecasters' Ranks based on the ATF and Accuracy Score

Persons	Answers	Day 1	Day 2	Day 3	Rank_{ATF}	AS	Rank_{AS}
Person 1	Yes	0.75	0.75	Yes	1	-0.5	1
	No	0.25	0.25				
	DBS	0.125	0.125				
Person 2	Yes		0.75	Yes	2	-0.25	2
	No		0.25				
	DBS		0.125				
Person 3	Yes	0.25	0.25	Yes	3	0.5	4
	No	0.75	0.75				
	DBS	1.125	1.125				
Person 4	Yes		0.25	Yes	4	0.25	3
	No		0.75				
	DBS		1.125				

Note. Answers = Possible future states of the world forecast by each person—a higher probability assigned to an answer means that a person deems this answer more likely than other answers; DBS = Daily Brier Score—the Brier Score calculated at the specific day when a forecast was active; AS = Accuracy Score by Cultivate; Day 1 = Probabilities that each person assigned to each possible future state of the world at day 1 (= probability forecasts at day 1); Day 2 = Probabilities that each person assigned to each possible future state of the world at day 2 (= probability forecasts at day 2); Day 3 = True future state of the world that was forecast; $\mathbf{Rank}_{\boldsymbol{ATF}}$= Expected rank of each person based on their predictions at day 1 and day 2 if the quality of forecasts is conceptualized using the aspects accuracy and timing; $\mathbf{Rank}_{\boldsymbol{AS}}$ = Rank based on the Accuracy Score by Cultivate.

The example given in Table 3 illustrates the critical issue with the *Accuracy Score*: Even though the *Accuracy Score* and rank of person 1 are better than the *Accuracy Score* and rank of

[2] The *Accuracy Score* takes the timing into account by the factor $1/T_q$, which rewards good forecasts that are made earlier than other good forecasts.

[3] Initially, we thought that the *Accuracy Score* would incorporate the timing of forecasts adequately. However, when we tried to understand how exactly the *Accuracy Score* works calculating the score on data we had collected during prediction tournaments we ran on two different Cultivate platforms in Fall 2019, we discovered that the *Accuracy Score* only takes the timing of good forecasts into account, whereas it does not take the timing of bad forecasts into account.

person 2, the *Accuracy Score* and rank of person 3 are worse than the *Accuracy Score* and rank of person 4. However, under the assumption of the increasing availability of relevant information over time, person 3 should be ranked higher than person 4. That is, the *Accuracy Score* does not adequately consider the timing of bad scores and therefore, is not an adequate operationalization of the quality of forecasts as conceptualized above.

It is important to note that we do not question the value of previous work building on these scores. On the contrary, previous research was mostly interested in the *accuracy* of forecasts and provided valuable insights into what determines the accuracy of forecasts. However, because not only the accuracy but also the timing of forecasts is relevant for business, we argue that future research should examine the quality of forecasts as conceptualized in this research note, which represents an integration of the concepts of accuracy and timing. As none of the existing scores seems to adequately operationalize the quality of forecasts, a new score that allows researchers to measure the quality of forecasts is needed. In the following, we are going to propose the *Quality Score* as a basis for research on the quality of forecasts.

The Quality Score. The *Quality Score* is formally defined as

$$QS_{iq} = \frac{\sum_{t=1}^{T_q} \sqrt{t} \sum_{c=1}^{C} (f_{iqtc} - o_{qc})^2}{\sum_{t=1}^{T_q} \sqrt{t}}$$

$$with \sum_{c=1}^{C} (f_{iqtc} - o_{qc})^2 = 2 \; if \; person \; i \; made \; no \; forecast \; on \; question \; q \; at \; time \; t,$$

where QS_{iq} is the *Quality Score* of person i on question q. Like the *Brier Score*, the *Quality Score* is a mean squared error measuring the deviation of the forecasts f_{iqt} of person i on question q at time t from the true state of the world o_{qc} when question q is resolved. But unlike the *Brier Score*, the *Quality Score* is a weighted arithmetic mean, because it weights the forecast error

$\sum_{c=1}^{C}(f_{iqtc} - o_{qc})^2$ depending on the timing of person i (i.e., when person i made their forecast) by multiplying the forecast error with the square root of the current value of t when person i made their forecast. The square root term $\sqrt{t}$ reflects the assumption of the increasing availability of relevant information over time. Thereby, the *Quality Score* weights the forecast error $\sum_{c=1}^{C}(f_{iqtc} - o_{qc})^2$ of person i on question q at time t heavier, the closer person i made their forecast relative to the realization of the future state of the world. That is, the *Quality Score* considers both the accuracy and the timing of the forecasts and indicates how accurately and timely person i forecast the future state of the world on average. Consequently, the *Quality Score* represents an adequate operationalization of the quality of forecasts as conceptualized above.

This corresponds to the logic of the *Brier Score* as the *Brier Score* is a measure of the forecasting error of a person. Under the assumption of the increasing availability of relevant information over time, it should be easier to forecast the future state of the world for a person the closer the person is to the realization of the future state of the world. This means that the errors of a person are more serious the closer the person is to the realization of the future state of the world. Therefore, the *Quality Score* weights errors that are made more closer to the realization of the future state of the world more heavily.

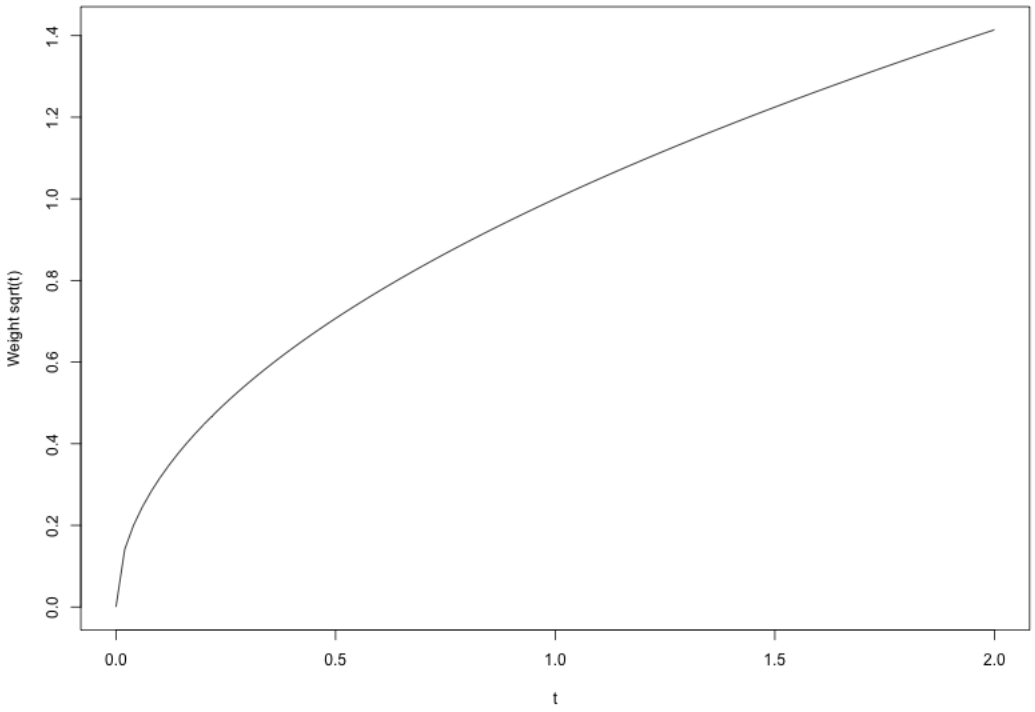

Figure 1. Square root function that is used to weight the errors of a person depending on when this person made their forecast—that is, depending on the value of t—as described in the *Quality Score* above. In our example, the realization of the future event occurs after day 2 (i.e., after $t = 2$). Consequently, errors made at day 2 ($t = 2$) are more heavily weighted than errors at day 1 ($t = 1$).

Furthermore, person i is assigned the largest possible forecast error (i.e., 2) on question q at time t if person i made no forecast on question q at time t. As persons can improve their forecast error by simply making a forecast, this strongly incentivizes persons to make their forecasts early. At the same time the *Quality Score* remains consistently interpretable as it can assume values between 0 and 2, where values closer to 0 indicate better forecasts; values of 0.5 are neutral; and values closer to 2 indicate worse forecasts.

The *Quality Score* adequately considers the accuracy and timing of both good and bad forecasts as illustrated in Table 4: the ranks of persons 1–4 correspond to the ranks of persons 1–4 under the assumption of the increasing availability of relevant information over time. For this

reason, we propose that Cultivate Labs should consider implementing a *Quality Score* instead of the currently used *Accuracy Score*. Without such a change, the suggestion made, for example, on the GJopen training video that "We want to reward early accurate forecasting" (https://goodjudgment.io/Training/KeepingScore/index.html) is not fully implemented in our view as late bad forecast are not scored worse than early bad forecasts.

Table 4: Comparison of the Forecasters' Ranks based on the ATF and Quality Score

Persons	Answers	Day 1	Day 2	Day 3	Rank_{ATF}	QS	Rank_{QS}
Person 1	Yes	0.75	0.75	Yes	1	0.125	1
	No	0.25	0.25				
	DBS	0.125	0.125				
Person 2	Yes		0.75	Yes	2	0.902	2
	No		0.25				
	DBS		0.125				
Person 3	Yes	0.25	0.25	Yes	3	1.125	3
	No	0.75	0.75				
	DBS	1.125	1.125				
Person 4	Yes		0.25	Yes	4	1.487	4
	No		0.75				
	DBS		1.125				

Note. Answers = Possible future states of the world forecast by each person—a higher probability assigned to an answer means that a person deems this answer more likely than other answers; DBS = Daily Brier Score—the Brier Score calculated at the specific day when a forecast was made; QS = Quality Score; Day 1 = Probabilities that each person assigned to each possible future state of the world at day 1 (= probability forecasts at day 1); Day 2 = Probabilities that each person assigned to each possible future state of the world at day 2 (= probability forecasts at day 2); Day 3 = True future state of the world that was forecast; $\mathbf{Rank}_{\boldsymbol{ATF}}$= Expected rank of each person based on their predictions at day 1 and day 2 if the quality of forecasts is conceptualized using the aspects accuracy and timing; $\mathbf{Rank}_{\boldsymbol{QS}}$ = Rank based on the Quality Score.

We now switch to another important topic for research of forecasting skills. When researchers conduct prediction tournaments, they may not only be interested in operationalizing the quality of forecasts but also in identifying so-called *superforecasters* (Schoemaker & Tetlock, 2016; Tetlock & Gardner, 2016)—persons with superior forecasting skills who consistently make better forecasts than other persons. To identify superforecasters, researchers frequently fall back

upon the scores described above, and thus implicitly assume that a better score does not only indicate the higher quality of a person's forecast, but also their higher forecasting skill. That is, the *Brier Score, Accuracy Score,* or *Quality Score* are implicitly assumed to be measures for a person's forecasting skill.

Although these scores represent reasonable measures for the forecasting skills of persons, these scores require that all persons forecast all questions in a prediction tournament (Merkle et al., 2017). If persons do not forecast all questions in a prediction tournament, differences in their scores may not be fully attributable to differences in their forecasting skills, as the questions in a prediction tournament likely differ in their characteristics. For example, the questions in a prediction tournament likely differ in their easiness. Predicting industry sales for the next year may be easier for industries with stable demands, such as utilities, grocery and health care, than predicting industry sales for industries with volatile demands, such as consumer discretionary products and the energy sector. Therefore, if persons do not forecast all questions in a prediction tournament, some may only have better scores because they selectively forecast easier questions––and not because they have better forecasting skills.[4] To repeat, only if all forecasters answer all questions in predictions tournament do the *Accuracy Score* or the *Quality Score* represent a reliable measure of forecasting skill.

As persons rarely forecast all questions in a prediction tournament, researchers frequently have to deal with the issue that differences in the scores may not be fully attributable to differences in their forecasting skills. To ensure that the scores measure the forecasting skills of persons,

[4] The *Accuracy Score* partly controls for differences in the easiness of questions because the crowd forecast is likely to be more accurate with easy questions and less accurate with hard questions, and the score of a forecaster is always calculated in relationship to how much better or worse the forecaster is in relationship to the crowd. Nonetheless, questions often still differ in their means, which indicates that they still differ in their easiness. For these reasons, differences in the scores of persons may still not be fully attributable to differences in their forecasting skills.

Kapoor and Wilde (2020), for example, control for differences in the characteristics of the forecast questions, such as their easiness, by including fixed question effects in their regression analyses. However, instead of using scores to measure the forecasting skills of persons, researchers can also use Item Response Theory (IRT) models, as IRT models allow researchers to disentangle forecasting skills from question characteristics (Merkle et al., 2017). IRT models thus provide an alternative way of operationalizing forecasting skill. In the following, we are going to describe the rationale of IRT models and refine a previous IRT model to offer an implementation of the conceptualization of the quality of forecasts as described in this research note.

Item Response Theory Models. IRT models are frequently used in psychology to model the responses of persons on a set of questions. The core idea of IRT models is that responses on a set of questions do not only depend on the characteristics of persons, but also on the characteristics of the questions. By modelling the influence of both person and question characteristics on the responses separately, IRT models allow researchers to disentangle the influence of person and question characteristics on the responses of persons.

It is important to note that IRT models are usually not used to model and test hypotheses. Instead, IRT models are frequently used to estimate person parameters, which—in the context of prediction tournaments—represent an estimate of each person's forecasting skill. That is, a person's forecasting skill is not measured by a score, such as the *Brier Score, Accuracy Score,* or *Quality Score*, but by the person parameters extracted from the IRT model. These person parameters can then be used in further statistical analyses.

IRT models have been used in the context of prediction tournaments before (Bo et al., 2017; Merkle et al., 2016, 2017). For example, Merkle et al. (2017) described an IRT model that models

the probability forecasts of persons on questions as a function of the questions' characteristics, the point in time when the forecast was made, and the persons' forecasting skills:

$$Y_{iq}^{*} \sim N\left(\mu_{iq}, \sigma_q^2\right)$$

$$\mu_{iq} = \beta_{0q} + \beta_{1q} d_{iq} + \lambda_q \theta_i$$

$$\sigma_q^2 = \xi_{0q},$$

where the probit-transformed probability forecast for the realized future state of the world Y_{iq}^{*} of person i on question q is assumed to follow a normal distribution $N\left(\mu_{iq}, \sigma_q^2\right)$, where μ_{iq} is the mean and σ_q^2 the variance.

μ_{iq} is modelled as a function of the person parameter θ_i representing the forecasting skill of person i. Researchers interested in the forecasting skills of persons can extract the estimates of all person parameters θ_i to obtain an estimate of the forecasting skill of each person i that can be used in further statistical analyses. The person parameters θ_i represent valid estimates of the forecasting skills of persons, because the IRT model controls for the effect of time β_{1q} on the easiness of question q—where the time variable d_{iq} refers to the number of time-units (usually days) the probability forecast for the probit-transformed probability forecast for the realized future state of the world Y_{iq}^{*} was away from the point in time when question q was resolved—and the characteristics of each question q—that is, question q's easiness β_{0q} and skill-utility[5] λ_q.

While question q's easiness β_{0q} is simply the mean probability forecast for the realized future state of the world on question q at time $d_{iq} = 0$, question q's skill-utility λ_q describes the effectiveness of the forecasting skill θ_i for making accurate and timely forecasts on question q. For example, the forecasting skill of a person may be effective for forecasting industry sales,

[5] What we call skill-utility is usually called discrimination in classical IRT modelling approaches.

whereas its effectiveness may be limited when predicting the oil price. Techniques constituting high forecasting skills, such as considering the base rate as a starting point for thinking about the probability of an event (Tetlock & Gardner, 2016), may simply work better for some questions than others. However, if the effectiveness of a person's forecasting skill varies across questions in a forecasting tournament, this means that some questions may be good indicators of a person's forecasting skill, whereas others may be bad indicators of a person's forecasting skill. Predicting industry sales accurately and timely should indicate higher forecasting skills than predicting the oil price accurately and timely, which has been shown to be radically unpredictable (Tetlock & Gardner, 2016, p. 270). Therefore, the skill-utility λ_q represents valuable information when researchers are interested in the forecasting skills of persons as it quantifies to what extent a question actually measures the forecasting skills of persons. IRT models allow researchers to control for these differences when estimating the persons' forecasting skills. σ_q^2 is assumed to vary across the questions because the variance of the probit-transformed probability forecasts for the realized future state of the world Y_{iq}^* likely differs between each question q.

IRT models can be construed as multilevel models consisting of two levels, where the level-1 model can be formally described as above with

$$Y_{iq}^* \sim N\left(\mu_{iq}, \sigma_q^2\right)$$

$$\mu_{iq} = \beta_{0q} + \beta_{1q} d_{iq} + \lambda_q \theta_i$$

$$\sigma_q^2 = \xi_{0q},$$

the level-1 model for the parameters varying across questions can be formally described with

$$\beta_{0q} = \beta_{00} + u_{0q}$$

$$\beta_{1q} = \beta_{10} + u_{1q}$$

$$\lambda_q = \lambda_0 + v_q$$

$$\xi_{0q} = \xi_{00} + w_{0q},$$

where the random effects are assumed to be multivariate normally distributed with

$$\begin{pmatrix} u_{0q} \\ u_{1q} \\ v_q \\ w_{0q} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{u0} & & & \\ \tau_{u0u1} & \tau_{u1} & & \\ \tau_{u0v} & \tau_{u1v} & \tau_v & \\ \tau_{u0w0} & \tau_{u1w0} & \tau_{vw0} & \tau_{w0} \end{pmatrix}\right],$$

and the level-1 model for the person-parameter varying across persons can be formally described with

$$\theta_i = \pi_0 + z_i,$$

where the random effect is assumed to be normally distributed with

$$z_i \sim N(0, \tau_z).$$

It is important to note that this IRT model corresponds to the conceptualization of the quality of forecasts as described in this research note: The IRT model does not only take the accuracy of the forecasts into account, but also the timing of the forecasts by including the effect of time β_{1q} in the model. The effect of time β_{1q} captures to what extent question q becomes easier over time, which can be regarded as a measure to what extent the availability of relevant information over time increases. Thereby, the assumption of the increasing availability of relevant information over time is modelled. Taken together, the person parameter θ_i represents an accurate measure of the forecasting skill of person i, because it considers both the accuracy and timing of the forecasts while the specific characteristics of each question, such as the questions' easiness and skill-utility, are controlled for.

Although Merkle et al.'s (2017) model represents an important step forward in measuring the forecasting skills of persons, its usefulness can be limited in practice as it only allows researchers to describe one probit-transformed probability forecast of a person per question. However, persons frequently make more than one probability forecast per question, because they

usually update their forecasts during the course of a prediction tournament. That is, Merkle et al.'s (2017) model does not allow researchers to capture changes in the probability forecasts of persons. To remedy this limitation, we propose a refined IRT model that builds on and extends the IRT model by Merkle et al. (2017): While we retain their approach to modelling both the accuracy and timing of the probability forecasts, we formulate a model that considers all probability forecasts of a person on a question during a prediction tournament.

Furthermore, we do not use the probit-transformed probability forecasts, because the normal distribution is not the ideal distribution for probability forecasts. The normal distribution is continuous and thus, adequate for unbounded continuous random variables ranging from $-\infty$ to $+\infty$. However, probability forecasts are bounded and range from 0 to 1. Although probability forecasts can be normalized through probit-transformation, this requires researchers to adjust the extreme forecasts 0 and 1, for example, to 0.001 and 0.999. Otherwise, the probit-transformed values of the extreme forecasts would not be interpretable as $probit(0) = -\infty$ and $probit(1) = \infty$, whereas the probit-transformed values of the adjusted extreme forecasts are interpretable as $probit(0.001) = -3.09$ and $probit(0.999) = 3.09$. That is, while the probit-transformation solves the issue that probability forecasts are not normally distributed, this requires researchers to adjust extreme forecasts, which entails some loss of information.

To avoid the probit-transformation of the extreme forecasts 0 and 1, we use the zero-one-inflated beta distribution as response distribution, because the zero-one-inflated beta distribution assigns each possible probability forecast, including the extreme forecasts 0 and 1, a sensible probability between 0 and 1. Even though the zero-one-inflated beta distribution is somewhat more complex than the normal distribution, the zero-one-inflated beta distribution is also more informative than the normal distribution. It captures important information about the probability

forecasts of persons through additional parameters, such as the zero-one-inflation probability (i.e., the probability that a 0 or 1 occurs) and the conditional one-inflation probability (i.e., the probability that a 1 rather than a 0 occurs). Most importantly, using the zero-one-inflated beta distribution allows researchers to predict probabilities instead of probits — that is, the scales of the predicted parameters correspond to the scale of the probability forecasts.

Formally, the first level of our IRT model can be written as

$$Y_{iqt} \sim ZOIB(\mu_{iqt}, \phi_q, \alpha_q, \gamma_q)$$

$$\mu_{iqt} = expit\left(\beta_{0q} + \beta_{1q} d_{iqt} + \lambda_q \theta_i\right)$$

$$\phi_q = \exp\left(\pi_{0q}\right)$$

$$\alpha_q = expit\left(\delta_{0q}\right)$$

$$\gamma_q = expit\left(\omega_{0q}\right),$$

where the probability forecast for the realized future state of the world Y_{iqt} of person i ($i = 1, 2, \dots, N$) on question q ($q \in 1, 2, \dots, Q$) at time t ($t = 1, 2, \dots, T_i$) is assumed to follow a zero-one-inflated beta distribution $ZOIB(\mu_{iqt}, \phi_q, \alpha_q, \gamma_q)$, where μ_{iqt} is the mean, ϕ_q is the precision, α_q is the zero-one-inflation probability, and γ_q is the conditional one-inflation probability.

μ_{iqt} is modelled as a function of the person parameter θ_i representing the forecasting skill of person i. Researchers interested in the forecasting skills of persons can extract the estimates of all person parameters θ_i to obtain an estimate of the forecasting skill of each person i that can be used in further statistical analyses. The person parameters θ_i represent valid estimates of the forecasting skills of persons, because the IRT model controls for the effect of time β_{1q} on the easiness of question q—where the time variable d_{iqt} refers to the number of time-units (usually days) the probability forecast for the realized future state of the world y_{iqt} was away from the point

in time when question q was resolved—and the characteristics of each question q—that is, question q's easiness β_{0q} and skill-utility λ_q.

The remaining parameters of the zero-one-inflated beta distribution—the precision ϕ_q, the zero-one-inflation probability α_q, and the conditional one-inflation probability γ_q—are assumed to vary across the questions as the precision, the probability that a 1 or 0 occurs, and the probability that a 1 rather than a 0 occurs of the probability forecasts likely differs between each question q. To link the linear combinations to the parameters of the zero-one-inflation beta distribution they predict, the corresponding inverse link functions of each parameter are used, which is the $expit(x) = \frac{\exp(x)}{1+\exp(x)}$ function and the natural exponential function $\exp(x)$. Thereby, the range of each linear combination is matched to the range of each parameter.

The level-2 model for the parameters varying across questions can be formally described with

$$\beta_{0q} = \beta_{00} + u_{0q}$$

$$\beta_{1q} = \beta_{10} + u_{1q}$$

$$\lambda_q = \lambda_0 + v_q$$

$$\pi_{0q} = \pi_{00} + p_{0q}$$

$$\delta_{0q} = \delta_{00} + q_{0q}$$

$$\omega_{0q} = \omega_{00} + r_{0q},$$

where the random effects are assumed to be multivariate normally distributed with

$$\begin{pmatrix} u_{0q} \\ u_{1q} \\ v_q \\ p_{0q} \\ q_{0q} \\ r_{0q} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{u0} & & & & & \\ \tau_{u0u1} & \tau_{u1} & & & & \\ \tau_{u0v} & \tau_{u1v} & \tau_{v} & & & \\ \tau_{u0p0} & \tau_{u1p0} & \tau_{vp0} & \tau_{p0} & & \\ \tau_{u0q0} & \tau_{u1q0} & \tau_{vq0} & \tau_{p0q0} & \tau_{q0} & \\ \tau_{u0r0} & \tau_{u1r0} & \tau_{vr0} & \tau_{p0r0} & \tau_{q0r0} & \tau_{r0} \end{pmatrix}\right],$$

and the level-2 model of our IRT model for the person-parameter varying across persons can be formally described with

$$\theta_i = \pi_0 + z_i,$$

where the random effect is assumed to be normally distributed with

$$z_i \sim N(0, \tau_z).$$

It is important to note that this IRT model, like the IRT model by Merkle et al. (2017), assumes the forecasting skill θ_i to be unidimensional. That is, each question is assumed to measure the same forecasting skill. If the assumption of unidimensionality is met, it becomes irrelevant whether a person forecast all questions in a prediction tournament or not as each question measures the same forecasting skill. Differences in the person parameters can then usually be fully attributed to differences in the persons' forecasting skills. Thereby, IRT models can provide a solution to the common issue that persons only rarely forecast all questions in a prediction tournament.

However, the assumption of unidimensionality can be violated in practice as it is possible that each question q requires a slightly different forecasting skill (Merkle et al., 2017).[6] For example, predicting industry sales for the next year in the finance & insurance sector may require a somewhat different forecasting skill than predicting industry sales for the next year in the automotive sector. If a person has more expert knowledge on the automotive industry than on the finance & insurance industry, this expert knowledge may be more helpful for making accurate and timely forecasts on questions related to the automotive industry than on questions related to the finance & insurance industry. That is, forecasting skill can often be construed as multidimensional, because each question measures a somewhat different forecasting skill.

[6] More specifically, if the assumption of unidimensionality is not met, the assumption of local independence is violated, which means that the residuals are correlated across the questions.

To model multidimensional forecasting skills, the model for the parameter μ_{iqt} can be adjusted by replacing the unidimensional person parameter θ_i with the multidimensional person parameter θ_{iq} representing the forecasting skill of person i on question q. Formally, we can write

$$\mu_{iqt} = expit\big(\beta_{0q} + \beta_{1q} d_{iqt} + \lambda_q \theta_{iq}\big).$$

That is, each person i is assigned q question-specific person parameters representing the question-specific forecasting skills of this person. Although this model assumes that each question measures a somewhat different forecasting skill, modelling multidimensional forecasting skills can nevertheless be useful as the estimates of the person parameters representing the forecasting skills of persons can be regarded as indicators of a person's overall forecasting ability. It is important to note that this IRT model also controls for question characteristics, such as the questions' easiness and skill-utility, which usually allows researchers to fully attribute differences in the question-specific person parameters to differences in the question-specific forecasting skills of persons.

Using the multidimensional person parameter θ_{iq} instead of the unidimensional person-parameter θ_i entails some changes to the level-2 model of the person-parameter. The person parameter θ_{iq} now varies both across persons and across questions, which can be formally described as

$$\theta_{iq} = \pi_0 + z_i + z_q,$$

where the random effect of the persons z_i is assumed to be normally distributed with

$$z_i \sim N(0, \tau_z)$$

and the random effect of the questions z_q is added to the level-2 model for the parameters varying across questions, where the random effects are now assumed to be multivariate normally distributed with

$$\begin{pmatrix} u_{0q} \\ u_{1q} \\ v_q \\ p_{0q} \\ q_{0q} \\ r_{0q} \\ z_q \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{u0} & & & & & & \\ \tau_{u0u1} & \tau_{u1} & & & & & \\ \tau_{u0v} & \tau_{u1v} & \tau_v & & & & \\ \tau_{u0p0} & \tau_{u1p0} & \tau_{vp0} & \tau_{p0} & & & \\ \tau_{u0q0} & \tau_{u1q0} & \tau_{vq0} & \tau_{p0q0} & \tau_{q0} & & \\ \tau_{u0r0} & \tau_{u1r0} & \tau_{vr0} & \tau_{p0r0} & \tau_{q0r0} & \tau_{r0} & \\ \tau_{u0z} & \tau_{u1z} & \tau_{vz} & \tau_{p0z} & \tau_{q0z} & \tau_{r0z} & \tau_z \end{pmatrix}\right].$$

Given the complexity of this IRT model, practitioners may wonder how such a complex IRT model can be fitted. Taking a Bayesian approach, the R package *brms* (Bürkner, 2017, 2020) proved to be powerful and reliable when we fitted this model to a simulated data set with 10,000 observations. *brms* allows users to define Bayesian models by writing standard syntax in R and fits these models using the Bayesian programming language Stan, which uses MCMC sampling via the adaptive Hamiltonian Monte Carlo (HMC) algorithm. The key advantage of HMC compared to other MCMC algorithms is that it works well with complex models. The models fitted in *brms* use minimally informative priors optimized for HMC by default.

It is important to note that this IRT model corresponds to the conceptualization of the quality of forecasts as described in this research note as the model considers both the accuracy and the timing of the probability forecasts. Furthermore, IRT models have several advantages over using the *Brier*, the *Accuracy* or our proposed *Quality Score* if researchers are interested in measuring the forecasting skills of persons: First, it becomes irrelevant whether a person forecast all questions in a prediction tournament or not if the assumption of unidimensionality is met, as in such a case each question measures the same forecasting skill. Thereby, IRT models can provide a solution to the common issue that persons only rarely forecast all questions in a prediction tournament. Furthermore, differences in the person parameters can usually be fully attributed to differences in forecasting skills, because IRT models allow researchers to control for question characteristics, such as the questions' easiness and skill-utility, and the effect of time on the

easiness of the questions. Second, even if the forecasting skills of persons are assumed to be multidimensional, IRT models are useful because estimates of the person parameters of IRT models are usually more accurate than the *Brier*, the *Accuracy* or our proposed *Quality Score* when it comes to measuring forecasting skills. The reason is that IRT models still allow researchers to control for question characteristics, such as the questions' easiness and skill-utility, and the effect of time on the easiness of the questions. Therefore, the multidimensional IRT model usually allows researchers to fully attribute differences in the question-specific person parameters to differences in question-specific forecasting skills, which can be regarded as indicators of a person's overall forecasting ability. Third, IRT models are more informative than the *Brier*, the *Accuracy* or our proposed *Quality Score*, because they provide estimates of each question's easiness, skill-utility, and the effect of time.

Moreover, our IRT model refines previous IRT models in three important ways: First, it is based on all probability forecasts of all persons on all questions at each point in time. That is, it is based on the maximum information available in prediction tournaments. Second, it models the probability forecasts of each person on all questions at each point in time by using the zero-one-inflated beta distribution representing an adequate probability distribution for probability forecasts ranging from 0 to 1. This allows researchers to model potentially interesting characteristics of the response distributions across questions, such as their precision, the probability that a 1 or 0 occurs, or the probability that a 1 rather than a 0 occurs. Third, it does not necessarily make the assumption of one single forecasting skill per person (assumption of unidimensionality) but allows users to model multiple question-specific forecasting skills for each person (assumption of multidimensionality).

Taken together, IRT models have several advantages over the *Brier*, the *Accuracy* or our proposed *Quality Score* when researchers are interested in measuring forecasting skills. Maybe the most important advantage is that IRT models usually allow researchers to attribute differences in the (question-specific) person parameters to differences in the persons' (question-specific) forecasting skills by controlling for question characteristics, such as the questions' easiness and skill-utility, and the effect of time on the easiness of the questions. Thereby, they allow researchers to measure forecasting skills accurately. Therefore, IRT models should be the preferred tool for measuring forecasting skills.

We hope that our research note will help advance efforts to create better measurement of forecasting skills both on forecasting platforms such as GJopen and the St. Gallen platform and also advance the use of IRT models in the context of prediction tournaments.

References

Bo, Y. E., Budescu, D. V., Lewis, C., Tetlock, P. E., & Mellers, B. (2017). An IRT forecasting model: Linking proper scoring rules to item response theory. *Judgment and Decision Making*, *12*(2), 90–103.

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C. (2020). Bayesian Item Response Modeling in R with brms and Stan. *ArXiv:1905.09501 [Stat]*. http://arxiv.org/abs/1905.09501

Kapoor, R., & Wilde, D. (2020). Peering Into a Crystal Ball: Foresight During Periods Of Industry Change. *Working Paper Version February 11,2020*.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., & Tetlock, P. (2015). Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. *Perspectives on Psychological Science*, *10*(3), 267–281. https://doi.org/10.1177/1745691615577794

Mellers, B., Tetlock, P., & Arkes, H. R. (2019). Forecasting tournaments, epistemic humility and attitude depolarization. *Cognition*, *188*, 19–26. https://doi.org/10.1016/j.cognition.2018.10.021

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & Tetlock, P. E. (2014). Psychological Strategies for Winning a Geopolitical Forecasting Tournament. *Psychological Science*, *25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255

Merkle, E. C., Steyvers, M., Mellers, B., & Tetlock, P. E. (2016). Item response models of probability judgments: Application to a geopolitical forecasting tournament. *Decision*, *3*(1), 1–19. https://doi.org/10.1037/dec0000032

Merkle, E. C., Steyvers, M., Mellers, B., & Tetlock, P. E. (2017). A neglected dimension of good forecasting judgment: The questions we choose also matter. *International Journal of Forecasting*, *33*(4), 817–832. https://doi.org/10.1016/j.ijforecast.2017.04.002

Schoemaker, P. J. H., & Tetlock, P. E. (2016). Superforccasting: How to Upgrade Your Company's Judgment. *Harvard Business Review*, *94*(5), 73–78.

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The Art and Science of Prediction*. Random House.