

# Accuracy of genomic breeding values revisited: Assessment of two established approaches and a novel one to determine the accuracy in two-step genomic prediction

G. Ni<sup>1,2</sup> | S. Kipp<sup>3</sup> | H. Simianer<sup>1</sup>  | M. Erbe<sup>1,4</sup> 

<sup>1</sup>Animal Breeding and Genetics Group, Georg-August-University, Göttingen, Germany

<sup>2</sup>School of Environmental and Rural Science, University of New England, Armidale, NSW, Australia

<sup>3</sup>Vereinigte Informationssysteme Tierhaltung w.V. (vit), Verden, Germany

<sup>4</sup>Institute for Animal Breeding, Bavarian State Research Centre for Agriculture, Grub, Germany

## Correspondence

M. Erbe, Institute for Animal Breeding, Bavarian State Research Centre for Agriculture, Grub, Germany.  
Email: malena.erbe@lfl.bayern.de

## Funding information

China Scholarship Council

## Summary

Selection decisions in genomic selection schemes are made based on genomic breeding values (GBV) of candidates. Thus, the accuracy of GBV is a relevant parameter, as it reflects the stability of prediction and the possibility that the GBV might change when more information becomes available. Accuracy of genomic prediction defined as the correlation between GBV and true breeding values (TBV), however, is difficult to assess, considering TBV of the candidates are not available in reality. In previous studies, several methods were proposed to assess the accuracy of GBV including methods using population parameters or parameters inferred from mixed-model equations. In practice, most approaches tended to overestimate the accuracy of genomic prediction. We thus tested approaches used in previous studies in order to assess the magnitude of bias. Analyses were performed based on simulated data under a variety of scenarios mimicking different livestock breeding programmes. Furthermore, we proposed a novel method and tested it both with simulated data and in a real Holstein data set. The new method provided a better prediction for the accuracy of GBV in the simulated scenarios.

## KEYWORDS

accuracy, genomic prediction, simulation

## 1 | INTRODUCTION

With the widespread availability of high-throughput single-nucleotide polymorphism (SNP) genotyping, genomic selection (GS) became widely used in livestock breeding (Hayes, Bowman, Chamberlain, & Goddard, 2009; Meuwissen, Hayes, & Goddard, 2013) and plant breeding (Jannink, Lorenz, & Iwata, 2010; Rincet et al., 2012), and displays dramatic advantages in genetic progress in both simulated (Habier, Fernando, & Garrick, 2013) and real (e.g., García-Ruiz et al., 2016) selection scenarios, especially for sex-limited traits or traits that can only be measured late in life (Meuwissen et al., 2013). Decisions

in GS schemes are made based on genomically estimated breeding values (GBV) of selection candidates. A GBV as used in this study is the prediction of an individual's true breeding value (TBV) derived from its SNP genotype and marker effects estimated in a set of genotyped and phenotyped animals of the same population. The accuracy of GBV, defined as the correlation between TBV and GBV, reflects the stability of the prediction and the possibility that the GBV might change when more information becomes available (Bijma, 2012). Furthermore, it is one of the key factors in the breeders' equation which is used to predict the expected response to selection (Falconer & Mackay, 1996).

In a two-step genomic prediction, pedigree-based breeding values (EBV) are first estimated with pedigree-based best linear unbiased prediction (BLUP), and then, GBV are estimated with EBV or their derivatives, for example, de-regressed proofs (DRP) or daughter yield deviations (DYD), used as quasi-phenotypes. To assess the predictive ability of GBV, the correlation between GBV and EBV in genomic prediction schemes can be easily calculated. The actually interesting correlation between GBV and TBV, namely the accuracy of genomic prediction, is hardly possible to assess, because TBV of the candidates are not available in reality.

In previous studies, several approaches have been suggested to assess the accuracy of GBV: one possibility is to use population and trait parameters (e.g., effective population size, size of chromosomes, heritability) to approximate the accuracy. The advantage of these approaches is that they can be used before data of selection candidates are collected (Wientjes, Veerkamp, & Calus, 2013). They give an overall assessment of the expected accuracy, which treats the studied samples as a whole but does not differentiate between animals with different amounts of information. Different authors suggested equations predicting accuracy from known population parameters (e.g., Daetwyler, Pong-Wong, Villanueva, & Woolliams, 2010; Goddard, Hayes, & Meuwissen, 2011) which, however, were shown to provide results which are not always matching the real values, especially when extrapolating predictions beyond the actually observed space (Erbe, Gredler, Seefried, Bapst, & Simianer, 2013).

The second class of approaches estimating the accuracy of GBV is using parameters inferred from the mixed-model equations (MMEs). Following Henderson (1975), accuracy of estimated breeding values for a given animal  $i$  ( $r_{BV_i}$ ) can be calculated from the prediction error variance for individual  $i$ ,  $PEV_i$ , which can be obtained from the inverse of the coefficient matrix of the MMEs, and the genetic variance  $var(A_i)$  as

$$r_{BV_i} = \sqrt{1 - \frac{PEV_i}{var(A_i)}}$$

with  $var(A_i) = A_{ii}var(A)$  and  $A_{ii}$  is the diagonal element of the numerator relationship matrix for individual  $i$ .

In principle, this type of assessment can also be used in genomic breeding value estimation implemented with GBLUP (VanRaden, 2008), where the pedigree-based numerator relationship matrix in BLUP is replaced by the genomic relationship matrix. However, this estimate of accuracy only holds under the absence of selection (Bijma, 2012; Dekkers, 1992; Gorjanc, Bijma, & Hickey, 2015) and tends to be biased when ignoring the changes in the (co)variance structure in selected populations (Dekkers,

1992; Henderson, 1975). An advantage of this approach is that a specific accuracy is obtained for each individual for which a breeding value is estimated.

A third way of assessing the accuracy of GBV is based on observed correlations between the TBV and different breeding value estimates as well as errors of estimates. One simple way is to use the correlation between EBV and GBV divided by the theoretical correlation between EBV and TBV ( $r_{GT} = r_{EG}/r_{ET}$ ). This formula or formulas replacing the EBV by corresponding derivatives thereof (e.g., DYD, or DRP) have been widely used in real data analysis (e.g., Hayes, Bowman et al., 2009; Luan et al., 2009; Saatchi et al., 2011) and are easy to implement, as  $r_{ET}$  (exactly or approximately) is available from MMEs of pedigree-based BLUP and  $r_{EG}$  can be empirically calculated as the correlation between EBV (DRP, DYD) and GBV. Note that  $r_{ET}$  from MME might be biased if there is selection and might need to be adjusted (e.g., Bijma, 2012). Amer and Banos (2010) suggested that the potential covariance between errors of GBV ( $\epsilon_G$ ) and errors of EBV ( $\epsilon_E$ ) needs to be taken into account when predicting the accuracy of GBV via this route. However,  $\epsilon_G$  and  $\epsilon_E$  usually are not available in real data so that their covariance is not known, and a formula including this parameter is not immediately applicable in practice. With the approaches described in this paragraph, the average accuracy in a given set of individuals can be studied.

In practical applications, most approaches were found to overestimate the accuracy of GBV (Goddard, 2009; Goddard et al., 2011; Hayes, Visscher, & Goddard, 2009). In addition, the magnitude of overestimation is unknown in real data sets, and little attention has been given to the quantification of how much these approaches overestimate the accuracy of GBV.

The first objective of this study was thus to test several of the approaches mentioned above with simulated data under a variety of scenarios mimicking different types of livestock breeding programmes (i.e., a basic scenario, a cattle-like and a pig-like one) and to measure the magnitude of overestimation. The second objective of this study was to suggest a novel and computationally feasible method that can provide a better prediction for the accuracy of GBV in real data sets and to assess the quality of the new approximation with both simulated and real data.

## 2 | MATERIAL AND METHODS

### 2.1 | Different approaches of estimating accuracy of genomic breeding values

All approximations of accuracies discussed in this section are thought to be calculated within a cohort of individuals

with comparable amount of information to obtain the average accuracy for this given set of individuals.

### 2.1.1 | Definition of accuracy and its relation to various previously suggested approaches of assessing the accuracy of GBV

In the following, we will use  $G$ ,  $E$  and  $T$  for the GBV, the EBV and the TBV, respectively. The correlation between GBV and EBV is defined as

$$r_{GE} = \frac{\text{cov}(\mathbf{G}, \mathbf{E})}{\sqrt{\text{var}(\mathbf{G})\text{var}(\mathbf{E})}} \quad (1)$$

Based on the theory of BLUP, we can assume  $\text{var}(\mathbf{G}) = \text{cov}(\mathbf{G}, \mathbf{T})$  and  $\text{var}(\mathbf{E}) = \text{cov}(\mathbf{E}, \mathbf{T})$ . If we assume  $\mathbf{G}$  and  $\mathbf{E}$  to be regressed sums of  $\mathbf{T}$  and a deviation ( $\boldsymbol{\varepsilon}_G$  and  $\boldsymbol{\varepsilon}_E$ , respectively) that comprise all effects except the additive genetic, we can write as  $\mathbf{G} = r_{GT}^2(\mathbf{T} + \boldsymbol{\varepsilon}_G)$  and  $\mathbf{E} = r_{ET}^2(\mathbf{T} + \boldsymbol{\varepsilon}_E)$ . Thus, Equation (1) can be expressed as follows:

$$r_{GE} = \frac{\text{cov}(r_{GT}^2(\mathbf{T} + \boldsymbol{\varepsilon}_G), r_{ET}^2(\mathbf{T} + \boldsymbol{\varepsilon}_E))}{\sqrt{\text{cov}(r_{GT}^2(\mathbf{T} + \boldsymbol{\varepsilon}_G), \mathbf{T})\text{cov}(r_{ET}^2(\mathbf{T} + \boldsymbol{\varepsilon}_E), \mathbf{T})}} \\ = \frac{r_{GT}^2 r_{ET}^2 (\text{var}(\mathbf{T}) + \text{cov}(\boldsymbol{\varepsilon}_G, \boldsymbol{\varepsilon}_E) + \text{cov}(\mathbf{T}, \boldsymbol{\varepsilon}_G) + \text{cov}(\mathbf{T}, \boldsymbol{\varepsilon}_E))}{r_{GT} r_{ET} \sqrt{(\text{var}(\mathbf{T}) + \text{cov}(\mathbf{T}, \boldsymbol{\varepsilon}_G))(\text{var}(\mathbf{T}) + \text{cov}(\mathbf{T}, \boldsymbol{\varepsilon}_E))}} \quad (2)$$

By rearranging Formula (2), we get

$$r_{GT} = \frac{r_{GE}}{r_{ET} \frac{(\text{var}(\mathbf{T}) + \text{cov}(\boldsymbol{\varepsilon}_G, \boldsymbol{\varepsilon}_E) + \text{cov}(\mathbf{T}, \boldsymbol{\varepsilon}_G) + \text{cov}(\mathbf{T}, \boldsymbol{\varepsilon}_E))}{\sqrt{(\text{var}(\mathbf{T}) + \text{cov}(\mathbf{T}, \boldsymbol{\varepsilon}_G))(\text{var}(\mathbf{T}) + \text{cov}(\mathbf{T}, \boldsymbol{\varepsilon}_E))}}} \quad (3)$$

The covariance between  $\mathbf{T}$  and  $\boldsymbol{\varepsilon}_E$  can be assumed to be zero since

$$\text{cov}(\mathbf{T}, \boldsymbol{\varepsilon}_E) = \text{cov}\left(\mathbf{T}, \frac{\mathbf{E}}{r_{ET}^2} - \mathbf{T}\right) = \frac{1}{r_{ET}^2} \text{cov}(\mathbf{T}, \mathbf{E}) - \text{var}(\mathbf{T}) \\ = \text{var}(\mathbf{T}) - \text{var}(\mathbf{T}) = 0$$

The same is true for  $\text{cov}(\mathbf{T}, \boldsymbol{\varepsilon}_G)$ .

Thus formula (3) simplifies to the following:

$$r_{GT} = \frac{r_{GE}}{r_{ET} \frac{(\text{var}(\mathbf{T}) + \text{cov}(\boldsymbol{\varepsilon}_G, \boldsymbol{\varepsilon}_E))}{\sqrt{\text{var}(\mathbf{T})\text{var}(\mathbf{T})}}} = \frac{r_{GE}}{r_{ET} \left(1 + \frac{\text{cov}(\boldsymbol{\varepsilon}_G, \boldsymbol{\varepsilon}_E)}{\text{var}(\mathbf{T})}\right)} \quad (4)$$

which is the approximation for predicting the accuracy of GBV suggested by Amer and Banos (2010) and will be denoted as “Acc\_AB” in the following. By further assuming that  $\boldsymbol{\varepsilon}_G$  and  $\boldsymbol{\varepsilon}_E$  are independent ( $\text{cov}(\boldsymbol{\varepsilon}_G, \boldsymbol{\varepsilon}_E) = 0$ ), we get

$$r_{GT} = \frac{r_{GE}}{r_{ET}} \quad (5)$$

which is the formula used in Hayes, Visscher et al. (2009) and will be referred as “Acc\_H” in the following.

### 2.1.2 | A novel approach for approximating the accuracy of GBV

A step-by-step derivation of the novel approach can be found in Appendix 1. The idea of derivation is to standardize the vectors of breeding values and to use properties of standardized vectors, for example, that correlation coefficients between them equal regression coefficients.

The derivation resulted in

$$r_{GT} \equiv r_{EG} r_{ET} + \omega \sqrt{(1 - r_{ET}^2)(1 - r_{EG}^2)} \quad (6)$$

which we will hereinafter denote as Acc\_N and in which  $\omega$  is a correlation (and thus bounded between  $-1$  and  $1$ ) of errors of regressions of standardized TBV on standardized EBV and standardized GBV, respectively, and can be interpreted as a weighting factor for the second term in the formula. In real data,  $\omega$  is an unknown quantity which has to be estimated somehow.

## 2.2 | Simulation of data

Accuracy of genomic prediction was evaluated for three alternative main simulated data sets with different population parameters, called basic, cattle-like and pig-like scheme. The simulation was performed using the software QMSim (Sargolzaei & Schenkel, 2009). Results will be presented for 20 independent replicates of the simulated data. Simulation will be shortly described in the following; further details regarding the differences between the schemes can be found in Figure S1.

### 2.2.1 | Genome

The simulated genome was the same for all schemes: the genome consisted of 10 chromosomes with 100 centiMorgan each. Initially, there were 3,000 polymorphic markers (SNPs) and 50 biallelic quantitative trait loci (QTLs) randomly distributed on each chromosome with an initial allele frequency of 0.5. No mutation was simulated. Markers and QTLs with a minor allele frequency (MAF)  $\geq 0.01$  in the last historical generation were selected and used in the simulation of the recent population. The additive allelic effects of QTLs were drawn from a gamma distribution with shape parameter 0.2. The positions of markers and QTLs across the genome were randomized in each of the 20 replications.

### 2.2.2 | Traits and selection

A quantitative trait with heritability of 0.2 or 0.5, respectively, was simulated. TBV were simulated by summing up all true additive QTL allelic effects. The simulation

included random selection of parents (abbreviated as “noSel”) or selection of males (in the basic and cattle-like scheme) or both parents (in the pig-like scheme) based on EBV with predefined accuracy (abbreviated as “Sel”) in each scheme, in which the predefined accuracy was calculated based on the available information in each scheme, for example, the heritability of the trait, and the number of progeny, as suggested in Falconer and Mackay (1996). EBV were simulated using QMSim. The phenotypes were obtained by adding random residual effects to TBV and simulated from generation 6 on for all individuals in the basic scheme and all females in the cattle-like and pig-like scheme.

### 2.2.3 | Simulation schemes

All generations, historical and recent ones, were discrete. For all schemes, 1,000 historical generations (for further details see Figure S1) were modelled. From the last historical generation, 500 males/1,000 females (500/10,000, 500/500) were randomly chosen to act as founders of the recent population in the basic (cattle-like, pig-like) scheme. In the following twelve generations, 500 males were mated to 1,000 dams that produced two offspring in the basic scheme. To mimic a real cattle breeding scheme, 500 sires were mated to 10,000 dams per generation in the cattle-like scheme. Each dam produced two progeny with a probability of 50% for male progeny resulting in each sire having 20 female offspring and 20 male offspring. For the pig-like scenario, 500 males were mated to 500 dams and each litter consisted of four pigs (two males and two females). Note that we only simulated the individuals used in the breeding scheme, but no production animals in the pig-like scheme.

### 2.3 | Estimation of pedigree-based and genomic breeding values in a two-step model for further analyses

In the first step, EBV of individuals were estimated based on the following animal model:

$$y_1 = \mathbf{1}\mu_1 + \mathbf{Z}\mathbf{a} + \mathbf{e}_1$$

where  $y_1$  is the vector of phenotypic records,  $\mu_1$  is the overall mean,  $\mathbf{Z}$  is the design matrix of breeding values,  $\mathbf{a} \sim N(0, \mathbf{A}\sigma_a^2)$  is the vector of breeding values and  $\mathbf{e}_1$  is the vector of random errors following a normal distribution  $\mathbf{e}_1 \sim N(0, \mathbf{I}\sigma_{e_1}^2)$ .  $\mathbf{A}$  is the pedigree-based numerator relationship matrix. Based on this model, the vector of pedigree-based estimated breeding values (EBV) is obtained via BLUP.

In the second step, the GBV of sires were estimated based on the following model:

$$y_2 = \mathbf{1}\mu_2 + \mathbf{W}\mathbf{g} + \mathbf{e}_2$$

where  $y_2$  is the vector of quasi-phenotypes (EBV in this case) of sires in the training population,  $\mu_2$  is the overall mean,  $\mathbf{W}$  is the design matrix corresponding to  $\mathbf{g}$ , the vector of the animals' GBV which was assumed to be distributed  $\mathbf{g} \sim N(0, \mathbf{G}\sigma_g^2)$  and  $\mathbf{e}_2$  is the vector of random errors following a normal distribution  $\mathbf{e}_2 \sim N(0, \mathbf{I}\sigma_{e_2}^2)$ .  $\mathbf{G}$  is the genomic relationship matrix according to VanRaden (2007) and was built from the SNP markers.

EBV, GBV and corresponding variance components were estimated using ASReml 3.0 (Gilmour, Gogel, Cullis, & Thompson, 2009) for each scenario and each replicate.

### 2.4 | Training and validation sets for the genomic breeding value prediction

In real-life scenarios, there are two situations in GS schemes in which measures of accuracy are considered. The first one is when a validation study is performed within a given set of individuals for which the phenotypes of all individuals and/or their progeny have already been collected. For the genomic prediction model, the available data are split into training and validation set with information being masked for individuals in the validation set. The accuracy of genomic prediction then is assessed based on the correlation between predicted genomic breeding values (GBV) and pedigree-based breeding values of individuals in the validation set. Validation studies comprise both random cross-validation and forward prediction schemes, in which the youngest individuals comprise the validation set. The other situation, in which measures of accuracy are of interest, is the real challenge in GS, when the phenotypes of candidates and/or their progeny are not available at the time point the GBV are predicted. Both situations (validation study, abbreviated as “VAL” in the following, and candidate prediction, abbreviated as “CAND”) were investigated for each simulated scenario, and the accuracy of GBV of sires in the validation set or candidate set, respectively, obtained with different formulas was assessed for each scenario in each replicate separately. For “VAL,” we assumed pedigree and phenotypes to be known up to generation 12, that is, all individuals up to generation 11 had pedigree-based breeding values based on own and/or progeny information. The validation set for assessing genomic prediction accuracy comprised all sires from generation 11. For “CAND,” we used sires in generation 11 as candidates in the genomic prediction model and assumed phenotypic information to be available only up to generation 11 for pedigree-based breeding value estimation; thus, the

candidates had neither own nor progeny performance in the cattle- and pig-like schemes, while they had one own, but no progeny performance in the basic scheme.

## 2.5 | Analyses based on simulated data sets

One aim of this study was to show how the different formulas for assessing the accuracy of genomic prediction perform in different scenarios. For studying this, we calculated  $r_{EG}$ ,  $r_{ET}$  and  $r_{GT}$  as empirical correlations as well as the covariance between the errors of genomic and pedigree-based breeding value estimation ( $cov(\mathbf{e}_G, \mathbf{e}_E)$ ) and used it to calculate approximated accuracies with different formulas (Acc\_H, Acc\_AB, Acc\_N). As an empirical  $r_{ET}$  would not be available in real data, we also calculated the average theoretical value of the accuracy of pedigree-based breeding values based on the prediction error variance (ignoring inbreeding) and assessed the influence of using this approximation of  $r_{ET}$  in the formulas. We further studied properties of  $\omega$ , which is needed to calculate Acc\_N and is the only unknown parameter which remains to be determined in this formula given estimates of  $r_{EG}$  and  $r_{ET}$  are available. We assessed the range of the optimal  $\omega$  by minimizing the mean squared error (MSE) over the 20 replicates between the true accuracy of GBV, which is available in simulated data, and the approximation with the new formula of a certain scheme (basic, cattle- or pig-like) with specific heritability and selection/no selection. As  $\omega$  in Acc\_N is defined as a correlation (see Appendix 1),  $\omega$  can only take values between  $-1$  and  $+1$ . In other words, all possible MSE can be inspected when  $\omega$  is moved from  $-1$  to  $+1$  (with a step length of 0.02). The optimal  $\omega$  (i.e., the value leading to the smallest MSE) was used for all replicates of a specific scenario when calculating Acc\_N for comparison with the other formulas.

## 2.6 | Analyses based on real cattle data

As it was also of interest to assess the property of  $\omega$  of the new approach in real data, we had the need of a data set with the following characteristics: individuals in the data set should have a genomic and a pedigree-based breeding value which was predicted without own or progeny information of the individuals included in the model. Furthermore, they should have a second estimated breeding value based on as many progeny as possible that could be used as a proxy for the unknown TBV. These criteria could be met with a data set of progeny-proven Holstein cattle sires for which results (EBV, reliabilities of EBV, and GBV) from two routine breeding value runs from different time points (runs in 2010 and 2014) were provided by vit Verden (<http://www.vit.de/>). We chose only individuals for further analyses that were in the candidate set and had GBV

and EBV with  $r_{ET}^2 \leq 0.85$  in 2010 and were in the training set and had EBV with  $r_{ET}^2 \geq 0.95$  in 2014 (used as proxy for TBV and denoted as TBV<sub>E</sub> hereinafter). Traits studied included different production traits (milk yield (MY), fat yield (FY), protein yield (PY), fat percentage (FP), protein percentage (PP)) as well as somatic cell score (SCS) and body condition score (BCS). Besides the empirical correlation between EBV and GBV ( $r_{EG}$ ), the empirical correlation between TBV<sub>E</sub> and EBV ( $r_{ET-E}$ ), and TBV<sub>E</sub> and GBV ( $r_{GT-E}$ ) could be approximated as well with this data set. Consequently, the optimal  $\omega$  for Acc\_N could be approximated as  $\frac{r_{GT-E} - r_{EG}r_{ET-E}}{\sqrt{(1-r_{ET-E}^2)(1-r_{EG}^2)}}$ , which made it possible to study the robustness of the weighting factor for different traits.

## 3 | RESULTS

### 3.1 | Simulation

On average, there were 26,530 (26,624, 30,000) SNPs and 442 (435, 500) QTLs on the simulated chromosomes in the cattle-like (pig-like, basic) scheme that were still polymorphic ( $MAF > 0.01$ ) in the last historical generation. The MAF distribution based on the simulated genotypes of sires from generation 6 to generation 11 in the first replicate of the noSel scenario with a heritability of 0.5 within the cattle scheme is shown in Figure S2 and is similar to the distribution that can be seen in real data with common SNP arrays. The linkage disequilibrium (LD) distribution between SNPs whose distance was smaller than two centiMorgan (cM) is shown in Figure S3 a-c for different schemes for “NoSel” scenarios. There was no significant difference in the level of LD between “Sel” and “NoSel” (data not shown) scenarios in all schemes. Comparing schemes, there was no significant difference between pig-like and cattle-like schemes, which both had a higher LD level than the basic scheme.

### 3.2 | MSE of Acc\_N with different $\omega$ and the optimal $\omega$

Figure 1 shows MSE as average of the squares of the difference between the empirical correlation between GBV and TBV and the value of genomic accuracy predicted by Acc\_N with varying values for  $\omega$  and assuming  $r_{ET}$  to be empirically determined. Compared to the Sel scenarios, the curves of the noSel scenarios were slightly more flat. The curves of CAND scenarios were more convex than the curves of VAL scenarios (e.g., Sel\_CAND versus Sel\_VAL), which means that CAND scenarios were more sensitive to the choice of  $\omega$ . In CAND scenarios, the optimal  $\omega$  of cattle-like schemes were larger than those of pig-like and basic schemes; however, there were fewer differences

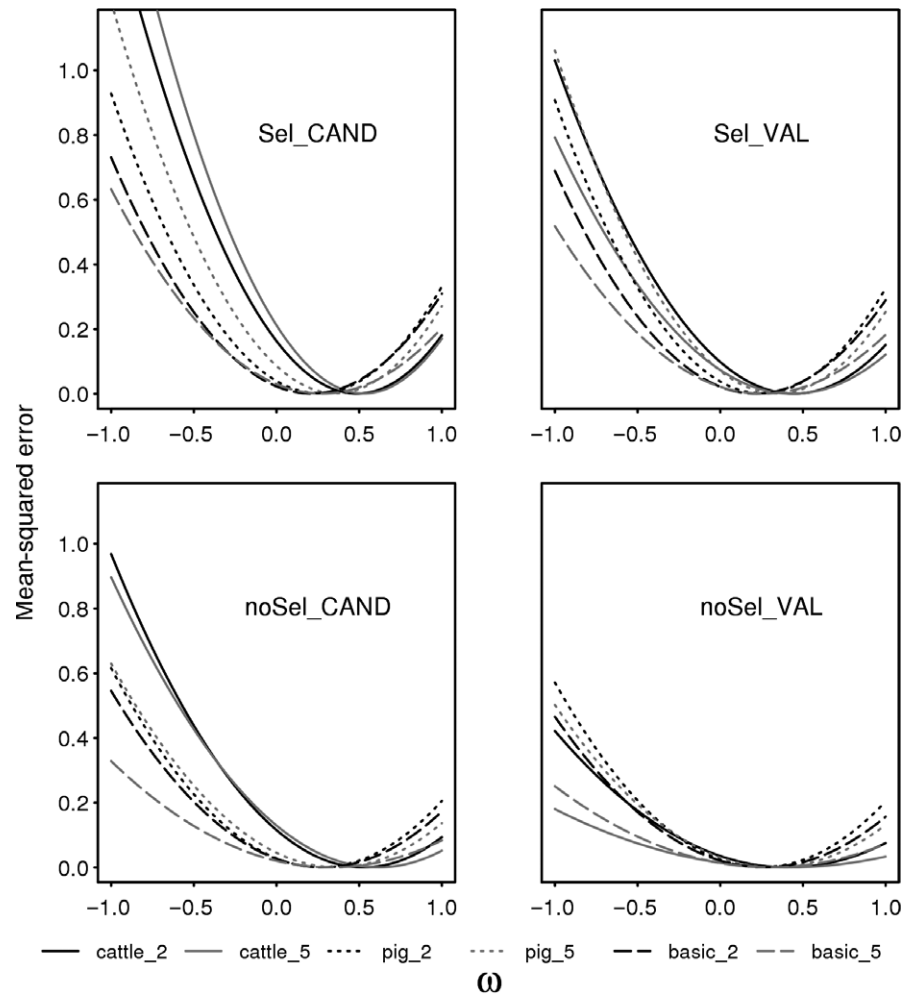


**FIGURE 1** Mean-squared errors (MSE) of the predicted accuracy of GBV calculated by Acc\_N in different simulation scenarios plotted against all possible values for  $\omega$  for basic, cattle-like and pig-like schemes with simulated heritabilities of 0.2 (denoted \_2) and 0.5 (denoted \_5)

in VAL scenarios. Furthermore, the optimal  $\omega$  (i.e., the one giving the minimum MSE over all replicates in a specific scenario) can be located and values are shown in Table 1. Optimal values were normally slightly greater in the Sel\_CAND scenarios compared to the noSel\_CAND scenarios, while there was no clear pattern in the VAL scenarios. Optimal  $\omega$  values (mean  $\pm$  SD) were relatively stable within all checked variants of the cattle-like ( $0.485 \pm 0.005$ ), pig-like ( $0.305 \pm 0.002$ ) and basic ( $0.268 \pm 0.001$ ) schemes, while means were clearly different when comparing across these three schemes, especially when comparing cattle-like schemes to the other two. Influence of using a theoretically derived  $r_{ET}$  value in the Acc\_N formula on optimal values of  $\omega$  was negligible in noSel scenarios, while it was more distinct in Sel scenarios (see Figure S4).

### 3.3 | Predicted accuracy with different approaches in different simulated scenarios

In Figures 2 and 3, predicted accuracies of GBV calculated from different approaches in VAL (2a and 3a) or CAND sets (2b and 3b) compared to the empirical correlation

between GBV and TBV are shown for scenarios with a heritability of 0.5. Results for scenarios with heritability of 0.2 are in Figure S5. In general, Acc\_N with optimal  $\omega$  showed considerably better performance than the other two approaches in most of the simulated cases. Acc\_H systematically overestimated the empirical accuracy in the simulated data sets, with a substantial proportion of estimates of accuracy larger than 1 and thus outside the parameter space, except for Sel\_VAL in the cattle-like scheme. The overestimation of Acc\_H was slightly stronger in the noSel scenarios than in Sel scenarios. Acc\_AB had a relative good performance in noSel scenarios, especially in the case when more information is available (i.e., noSel\_VAL), while it clearly underestimated the accuracy in many Sel scenarios. It should be noticed that when the empirical correlation between GBV and TBV was lower, the heritability was lower and/or the information was less (CAND in comparison with VAL), results obtained with Acc\_H and Acc\_AB had a stronger tendency to deviate from the empirical accuracy. The prediction with Acc\_N, however, was always in the same range regardless of the level of empirical accuracies of GBV in noSel scenarios. As

**TABLE 1** The optimal values for  $\omega$  in different scenarios<sup>a</sup>

|            | cattle_2 | cattle_5 | pig_2 | pig_5 | basic_2 | basic_5 |
|------------|----------|----------|-------|-------|---------|---------|
| Sel_CAND   | 0.49     | 0.53     | 0.25  | 0.36  | 0.21    | 0.28    |
| Sel_VAL    | 0.47     | 0.41     | 0.25  | 0.34  | 0.21    | 0.26    |
| noSel_CAND | 0.53     | 0.61     | 0.27  | 0.36  | 0.28    | 0.33    |
| noSel_VAL  | 0.41     | 0.40     | 0.25  | 0.33  | 0.27    | 0.30    |

<sup>a</sup>Scenarios differed in the selection scheme (basic, cattle-like, pig-like), the heritability of the simulated trait (0.2 denoted as \_2, or 0.5 denoted as \_5), the presence of selection ("Sel") or not ("noSel") as well as the information amount available in the background for the predicted individuals (like in a validation study, VAL, or in a real candidate situation, CAND).

empirical  $r_{ET}$  would not be available in praxis, we also checked the results obtained with Acc\_N, Acc\_H and Acc\_AB when  $r_{ET}$  was calculated based on mixed-model theory. Figure 4a–c show results for the VAL scenarios within the basic, cattle and pig schemes and a heritability of 0.5. It can clearly be seen that bias in noSel scenarios was less strong than in the Sel scenarios, especially in the cattle and the basic scheme. For both Acc\_H and Acc\_AB, prediction was better when a theoretically derived  $r_{ET}$  was used in the formulas. Acc\_N again turned out to be the most stable approach which is in part also due to the fact that the optimal  $\omega$  was calculated using the respective  $r_{ET}$  value (theoretically derived or empirical one) which was later also used in the formula to predict  $r_{GT}$ .

Changes in results when using a theoretically derived  $r_{ET}$  instead of an empirical  $r_{ET}$  were similar in tendency, but even more pronounced in scenarios with heritability of 0.2 and/or less information density for the EBV of the candidates like in the CAND scenarios (results not shown).

### 3.4 | Covariance between TBV, errors of GBV and errors of EBV

The degree of deviation of the approximated accuracy of GBV from the empirical  $r_{GT}$  by any of the considered methods is determined by omitting  $cov(\mathbf{T}, \mathbf{\epsilon}_E)$ ,  $cov(\mathbf{T}, \mathbf{\epsilon}_G)$ , and  $cov(\mathbf{\epsilon}_G, \mathbf{\epsilon}_E)$ , which are not known in reality, but available in simulated data (Figure 5a,b). Deviations from zero tended to be larger in CAND scenarios than in VAL scenarios. In Sel scenarios regardless of CAND or VAL,  $cov(\mathbf{T}, \mathbf{\epsilon}_E)$ ,  $cov(\mathbf{T}, \mathbf{\epsilon}_G)$  and  $cov(\mathbf{\epsilon}_G, \mathbf{\epsilon}_E)$  were all significantly different from zero. In noSel scenarios,  $cov(\mathbf{T}, \mathbf{\epsilon}_E)$  was found to be close to zero, which is in agreement with the assumption of traditional BLUP, while  $cov(\mathbf{T}, \mathbf{\epsilon}_G)$  deviated from zero in most scenarios. Values for  $cov(\mathbf{\epsilon}_G, \mathbf{\epsilon}_E)$  tended to be greater than zero in all Sel scenarios.

### 3.5 | Results from the real cattle data

A total of 1,271 individuals were chosen based on the two criteria described in *Material and Methods* to calculate the

correlations necessary to obtain an estimate for the optimal values for  $\omega$  (Table 2). Distributions of reliabilities of EBVs for MY, FY, PY and SCS are shown in Figure S6.

The optimal values for  $\omega$  were relatively stable for MY, PY, FY and SCS and ranged between 0.608 and 0.675. Compared to those traits, estimated optimal  $\omega$  were clearly higher for FP and PP and smaller for BCS.

## 4 | DISCUSSION

### 4.1 | Previous approaches

Accuracy of GBV is a critical parameter when performing GS, because it determines the accuracy of selection decisions and further is a critical parameter in optimization of the design of a breeding scheme.

For Acc\_AB and Acc\_H, specific assumptions have to be made that are related to BLUP properties:

1.  $cov(\mathbf{T}, \mathbf{\epsilon}_E) = cov(\mathbf{T}, \mathbf{\epsilon}_G) = 0$
2.  $var(\mathbf{G}) = cov(\mathbf{G}, \mathbf{T})$  and  $var(\mathbf{E}) = cov(\mathbf{E}, \mathbf{T})$

For Acc\_H, it is additionally assumed that  $cov(\mathbf{\epsilon}_E, \mathbf{\epsilon}_G) = 0$ . Based on our simulation results, assumption (1) was shown not to hold in many cases, especially when the data are from a population under selection. Figure S7 shows (for a specific scenario) that, while assumption (2) seems to hold for  $var(\mathbf{E})$  and  $cov(\mathbf{E}, \mathbf{T})$ ,  $var(\mathbf{G})$  was consistently smaller than  $cov(\mathbf{G}, \mathbf{T})$  even in noSel scenarios. This makes clear why we could not expect the two approaches Acc\_AB and Acc\_H to yield reliable results.

In previous studies, one popular approach of approximating the accuracy of GBV was Acc\_H (i.e.,  $r_{EG}/r_{ET}$ ), as it is easy to apply in real data sets by obtaining  $r_{EG}$  as the empirical correlation between EBV and GBV, and obtaining  $r_{ET}$  from the MMEs of traditional BLUP. In real cattle breeding schemes, particularly in validation studies in which there are bulls with information from many daughters, the accuracy of EBV is reasonably high (even close to one). Thus,

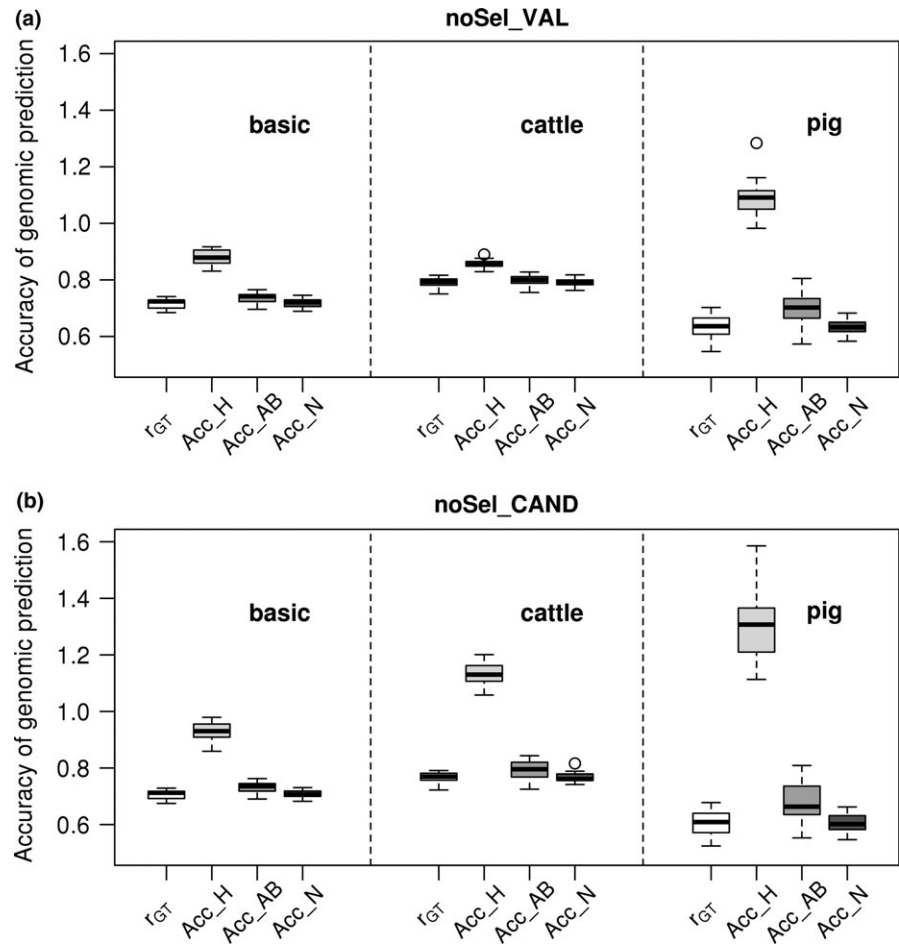


**FIGURE 2** (a–b) Boxplots of the empirical accuracy of GBV ( $r_{GT}$ ) and the accuracy of GBV estimated by different approaches in validation (VAL, (a) or candidate sets (CAND, (b) assuming no selection with a simulated heritability of 0.5 (20 replicates for each scenario). For Acc\_N, optimal values for  $\omega$  were used

dividing by  $r_{ET}$  does not have a strong effect on  $r_{EG}$ . Consequently, results obtained for Acc\_H were in a reasonable range in the simulated VAL cattle schemes with heritability of 0.5 (see Figures 2a and 3a). However, an overestimation of accuracy of GBV was still discovered frequently (Goddard, 2009; Goddard et al., 2011; Hayes, Visscher et al., 2009) in cattle and other species. Furthermore, the magnitude of inaccuracy of Acc\_H can hardly be quantified with real data. With less favourable data structures, Acc\_H appears to be substantially biased, partly resulting in estimated accuracies larger than 1 (e.g., Wellmann et al., 2013), which we also observed in some simulated scenarios and even in the real cattle data (results not shown).

Acc\_AB, calculated as  $\frac{r_{GE}}{r_{ET} \left( 1 + \frac{\text{cov}(\epsilon_G, \epsilon_E)}{\text{var}(T)} \right)}$ , conceptually is a relatively precise formula to estimate the accuracy of GBV in most cases. In the simulated data, we found that out of the three covariances  $\text{cov}(\epsilon_G, \epsilon_E)$ ,  $\text{cov}(\epsilon_E, T)$  and  $\text{cov}(\epsilon_G, T)$  that would have to be considered for an exact solution, the largest values were obtained for  $\text{cov}(\epsilon_G, \epsilon_E)$  on average which thus seems to be the most important covariance that must not be omitted. Acc\_AB nevertheless tended to underestimate accuracy in some selection scenarios when empirical  $r_{ET}$  was used. Results looked better with

theoretically derived  $r_{ET}$  used in the formula. Empirical and theoretically derived  $r_{ET}$  differ in selected populations (see e.g., Bijma, 2012), and thus, all other factors (including  $\epsilon_E$  as its calculation is based on empirical  $r_{ET}$  versus  $r_{ET}$  derived from PEV) also change. The upward-biased estimates of  $r_{ET}$  seem to counterbalance the missing  $\text{cov}(\epsilon_G, \epsilon_E)$  and thus let the result look better, but note that this enhancement may be a kind of an artefact. In any case, the non-availability of covariance between errors of GBV and errors of EBV limits the applicability of this approach in real data analyses.

In general, from this study we could see that deviations of results of these approaches Acc\_H and Acc\_AB from the empirical correlation between GBV and TBV tended to be higher when less information is available in the pedigree-based breeding value estimation (CAND versus VAL, cattle versus pig) or the genetic influence on the trait is smaller (heritability 0.2 versus 0.5, also leading to less reliable pedigree-based breeding values).

## 4.2 | Selection and estimates of $r_{ET}$

All approaches used in this study were based on estimates of  $r_{ET}$ . With the simulated data,  $r_{ET}$  derived from

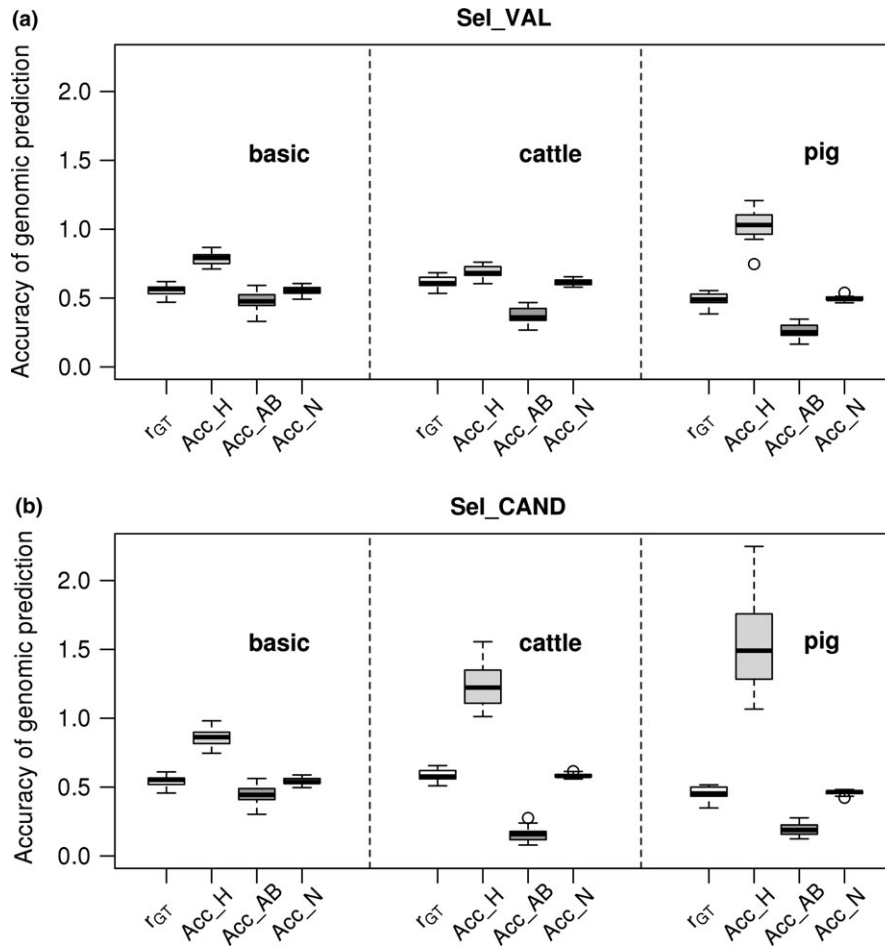


**FIGURE 3** (a–b) Boxplots of the empirical accuracy of GBV ( $r_{GT}$ ) and the accuracy of GBV estimated by different approaches in validation (VAL, (a) or candidate sets (CAND, (b) assuming selection with a simulated heritability of 0.5 (20 replicates for each scenario). For Acc\_N, optimal values for  $\omega$  were used

PEV and  $r_{ET}$  estimated as empirical correlation led to different results for all approaches, especially when selection took place. This is clear considering that selection influences the genetic variance, and thus, calculating  $r_{ET}$  using the estimate of  $var(T)$  reflecting the additive genetic variance in the unselected base population may lead to biased results. This problem has been addressed by different authors (e.g., Bijma, 2012; Gorjanc et al., 2015). Bijma (2012) defined formulas for estimating  $r_{ET}$  in case of selection which are applicable as long as selection intensities and truncation points of selection are known and selection schemes are not too complicated. Applied to our simulated data, bias could be reduced with Bijma's formulas in results of all approaches when  $r_{ET}$  derived from PEV (results not shown). One should, however, note that even with very accurate estimates of  $r_{ET}$  covariances between errors of estimation themselves and between TBV and errors of estimation may still occur causing bias in the estimates of prediction accuracy (especially in selected populations) in cases estimation approaches neglect them.

### 4.3 | Novel approach Acc\_N and weighting factor $\omega$

Our novel approach aims to describe the accuracy of genomic prediction as exact as possible, namely by equating

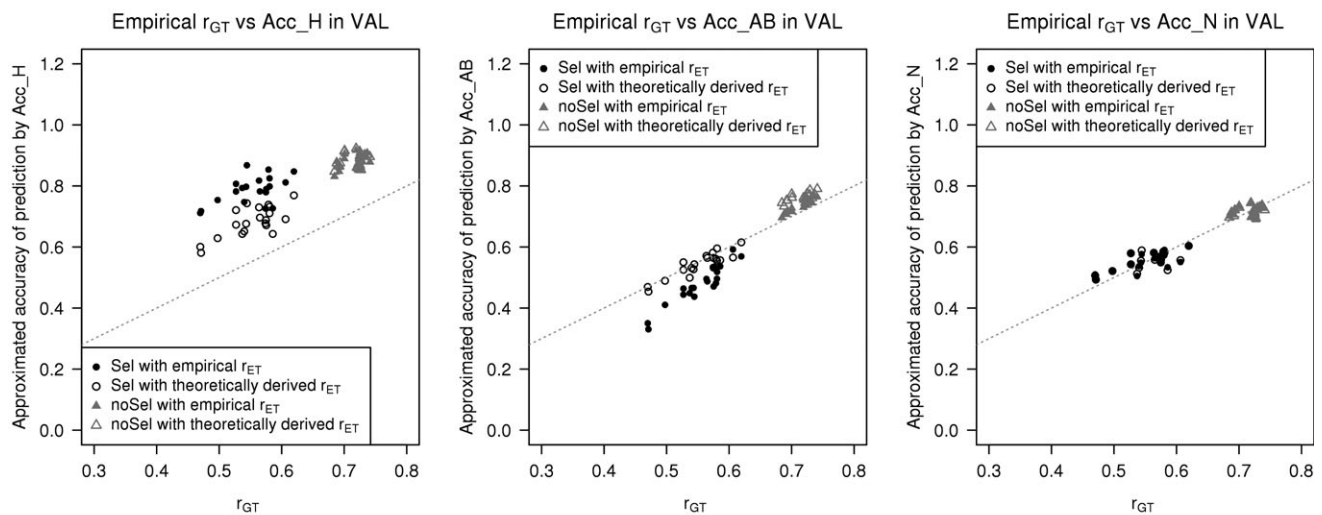
$$r_{GT} \equiv r_{EG}r_{ET} + \omega \sqrt{(1 - r_{ET}^2)(1 - r_{EG}^2)}$$

This equation provides a reliable, though under practical conditions still wide range for the true accuracy  $r_{GT}$  (see example below). As  $\omega$  is a correlation and thus is confined to the parameter space  $-1 \leq \omega \leq 1$ , the permissible space for  $r_{GT}$  is as follows:

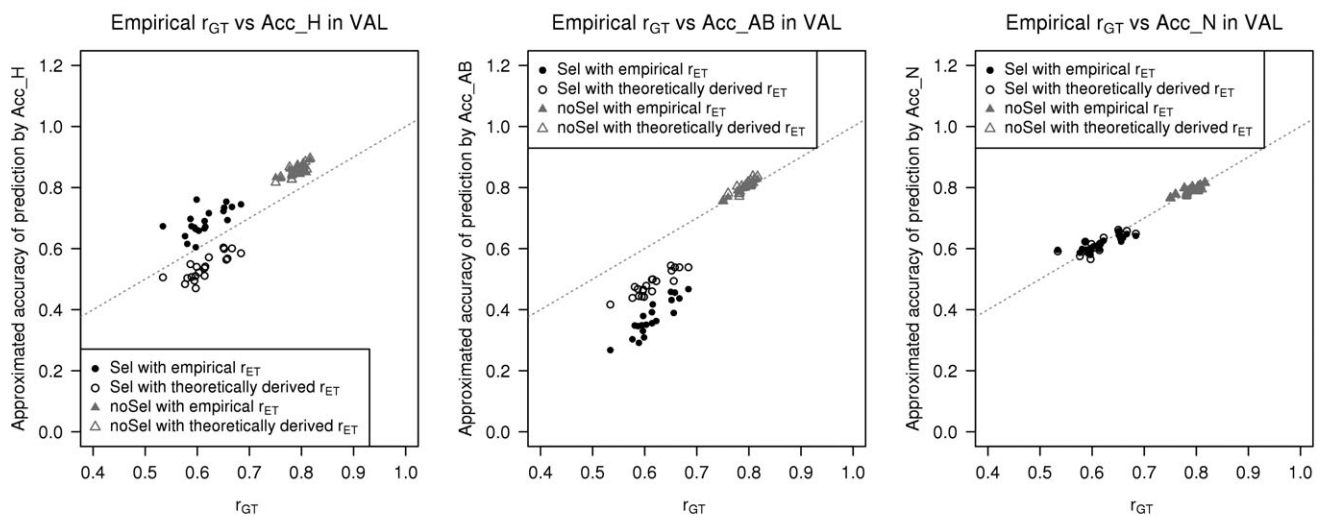
$$r_{EG}r_{ET} - \sqrt{(1 - r_{ET}^2)(1 - r_{EG}^2)} \leq r_{GT} \leq r_{EG}r_{ET} + \sqrt{(1 - r_{ET}^2)(1 - r_{EG}^2)}$$

One major advantage of this derivation is that it is not based on any assumptions related to statistical properties of BLUP. This means that if the empirical correlation between EBV and TBV and the optimal  $\omega$  was known, Acc\_N should provide the exact empirical accuracy of GBV.

### (a) Selection scheme: Basic



### (b) Selection scheme: Cattle like



### (c) Selection scheme: Pig like

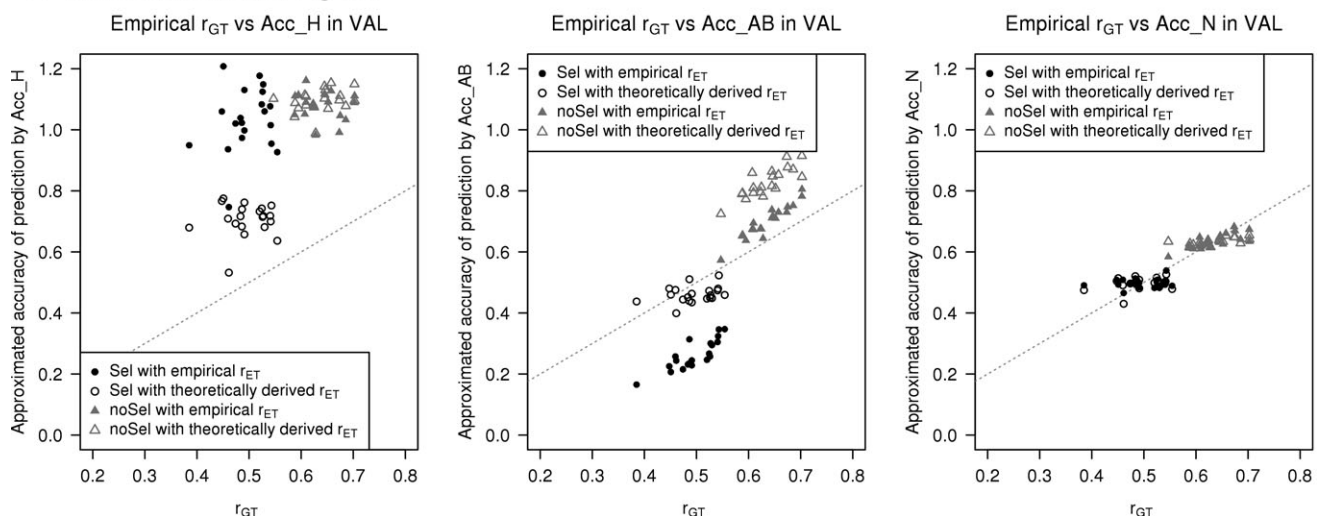


**FIGURE 4** (a–c) Comparison of empirical accuracy of GBV ( $r_{GT}$ ) and accuracy of GBV estimated by different approaches with accuracy of EBV ( $r_{ET}$ ) determined either empirically based on MME theory in all VAL scenarios in basic (a), cattle-like (b) or pig-like (c) schemes simulated with a heritability of 0.5 and either selection (Sel) or no selection (noSel). Twenty replicates were performed per scenario

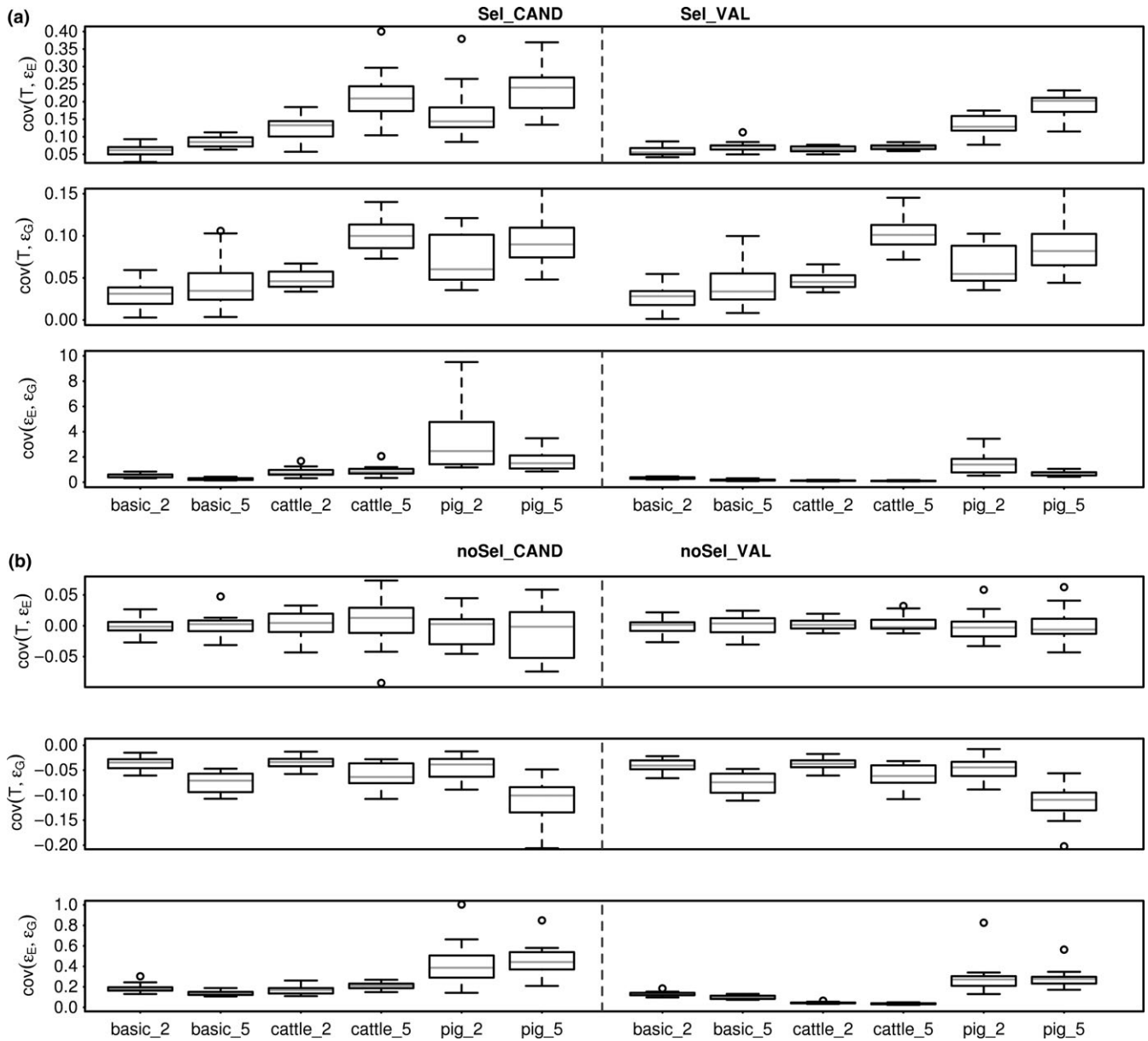


**FIGURE 5** (a–b) Covariance between errors of EBV and TBV [ $cov(T, \epsilon_E)$ ], covariance between errors of GBV and TBV [ $cov(T, \epsilon_G)$ ] and covariance of errors of EBV and errors of GBV [ $cov(\epsilon_E, \epsilon_G)$ ] in Sel\_CAND and Sel\_VAL (a) and noSel\_CAND and noSel\_VAL (b) scenarios

Acc\_N is computationally applicable and convenient in real data, provides smaller MSE for the predicted accuracy of GBV calculated and appears robust in various scenarios. Acc\_N also appeared to be working well with both empirical  $r_{ET}$  and theoretically derived  $r_{ET}$  as long as optimal  $\omega$  was estimated using  $r_{ET}$  in the way it was used later in the formula. The results of this study suggested that the new approach proposed here holds good potentials for estimating the accuracy of GBV in a more accurate way compared to two approaches used before.

As shown in Figure 1, with more information available, the curves of MSE for the predicted accuracy of GBV calculated by Acc\_N were more flat and less

sensitive to the choice of  $\omega$ . This might be because when more information is available, the accuracy of EBV (GBV) is higher and  $r_{EG}$  and  $r_{ET}$  are increased. Therefore,  $1 - r_{ET}^2$  and  $1 - r_{EG}^2$  in Equation (6) is reduced (tends to 0), causing the estimation being less sensitive to the choice of the weighting factor.

Acc\_N might also help to get some insight into another topic: The correlation between EBV (or its derivatives e.g., DRP, DYD) and GBV ( $r_{EG}$ ) is a widely used measure to assess quality of prediction especially in cross-validation scenarios (Luan et al., 2009), although it is not claimed to be an approximation of the accuracy of GBV in the strict sense. Equation (6) provides the link between  $r_{EG}$  and  $r_{GT}$ :

**TABLE 2** The optimal values for  $\omega$  and correlations between EBV, GBV and TBV<sub>E</sub> for the real cattle data<sup>a</sup>

| Traits               | $\omega$ | $r_{ET-E}$ | $r_{GT-E}$ | $r_{EG}$ |
|----------------------|----------|------------|------------|----------|
| Milk yield           | 0.670    | 0.526      | 0.768      | 0.572    |
| Fat yield            | 0.656    | 0.520      | 0.757      | 0.573    |
| Protein yield        | 0.608    | 0.518      | 0.723      | 0.571    |
| Fat percentage       | 0.798    | 0.496      | 0.850      | 0.528    |
| Protein percentage   | 0.766    | 0.567      | 0.846      | 0.612    |
| Somatic cell score   | 0.675    | 0.467      | 0.753      | 0.515    |
| Body condition score | 0.529    | 0.455      | 0.623      | 0.445    |

<sup>a</sup> $r_{ET-E}$  is the correlation between the EBV from 2010 and the EBV in 2014 assumed to be a good estimate of the TBV.  $r_{GT-E}$  is the correlation between the GBV from 2010 and the EBV assumed to be a good estimate of the TBV in 2014.  $r_{EG}$  is the correlation between the EBV and GBV from 2010.  $\omega$  is the parameter for Acc<sub>N</sub> estimated from the available data.

as  $-1 \leq \omega \leq 1$ , the range of  $r_{GT}$  is between  $r_{EG}r_{ET} + \sqrt{(1-r_{EG}^2)(1-r_{GT}^2)}$  and  $r_{EG}r_{ET} - \sqrt{(1-r_{EG}^2)(1-r_{GT}^2)}$ . This means that if you obtain estimates of for example  $r_{EG} = 0.7$  and  $r_{ET} = 0.8$ , the underlying  $r_{GT}$  can still be between 0.13 and 0.99. This makes clear that while  $r_{EG}$  is valuable to be used for model comparison or comparing results across studies (Daetwyler, Calus, Pong-Wong, de los Campos, & Hickey, 2013), this parameter is shown to be not robust as a proxy for the accuracy of GBV.

In this study, when applying GBLUP, EBV were used as quasi-phenotype to estimate GBV for the respective traits. However, progeny performances are often used as quasi-phenotype in reality as well. For example, in cattle breeding schemes, DYD are commonly employed, especially in the international genetic evaluation of bulls (Liu, Reinhardt, Bünger, & Reents, 2004; Mrode & Swanson, 2004; Szyda, Ptak, Komisarek, & Żarnecki, 2008), where DYD is defined as the weighted average of a bull's daughter yields corrected for all fixed effects. Thus, we estimated prediction accuracy of GBV with Acc<sub>N</sub> using DYD as phenotypic data in one of the simulated scenarios (cattle\_2\_Sel\_VAL) to test the variety of weighting factors. The mean ( $\pm$ SD) of the optimal weighting factor was 0.47 ( $\pm$ 0.04) when using DYD as input, compared to 0.45 ( $\pm$ 0.04) using EBV as input when using  $r_{ET}$  derived from PEV in the formulas. The correlation between weighting factors when using DYDs versus EBVs as input was 0.87.

The question is still on how to determine  $\omega$  in practice. For this, we suggested an approach to approximate an empirical value from a suitable data set reflecting two separate time points. It has to be studied further whether weighting factors derived with this strategy will be optimal in independent data sets for the same trait or could even be used across traits with for example similar genomic architecture or population parameters. Moreover, larger studies

are needed to figure out how stable  $\omega$  estimates are within trait when the amount of information or population structure changes and how  $\omega$  can be approximated when the suggested approach cannot be implemented. The latter can be true when working with for example low-heritability traits or new traits with limited number of available phenotypes so that highly reliable EBV are not available.

## ACKNOWLEDGEMENTS

G.N. acknowledges financial support from China Scholarship Council. The authors are grateful for friendship with and inspiration through Daniel Gianola, one of the most influential animal breeding scientists of the last decades. This work was in part motivated by intensive discussions with him during repeated research stays at Göttingen.

## REFERENCES

- Amer, P. R., & Banos, G. (2010). Implications of avoiding overlap between training and testing data sets when evaluating genomic predictions of genetic merit. *Journal of Dairy Science*, 93, 3320–3330.
- Bijma, P. (2012). Accuracies of estimated breeding values from ordinary genetic evaluations do not reflect the correlation between true and estimated breeding values in selected populations. *Journal of Animal Breeding and Genetics*, 129, 345–358.
- Daetwyler, H.D., Calus, M.P.L., Pong-Wong, R., de los Campos, G., & Hickey, J.M. (2013). Genomic prediction in animals and plants: Simulation of data, validation, reporting, and benchmarking. *Genetics*, 193, 347–365.
- Daetwyler, H. D., Pong-Wong, R., Villanueva, B., & Woolliams, J. A. (2010). The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. *Genetics*, 185, 1021–1031.
- Dekkers, J. C. M. (1992). Asymptotic response to selection on best linear unbiased predictors of breeding values. *Animal Production*, 54, 351–360.
- Erbe, M., Gredler, B., Seefried, F. R., Bapst, B., & Simianer, H. (2013). A function accounting for training set size and marker density to model the average accuracy of genomic prediction. *PLoS ONE*, 8, e81046.
- Falconer, D.S., & Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics*. Harlow, Essex, UK: Longmans Green.
- García-Ruiz, A., Cole, J. B., VanRaden, P. M., Wiggans, G. R., Ruiz-López, F. J., & Van Tassell, C. P. (2016). Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. *Proceedings of the National Academy of Sciences of the United States of America*, 113, E3995–E4004.
- Gilmour, A.R., Gogel, B.J., Cullis, B.R., & Thompson, R. (2009). *ASReml user guide*. Hemel Hempstead, UK: VSN International Ltd.
- Goddard, M. (2009). Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica*, 136, 245–257.
- Goddard, M. E., Hayes, B. J., & Meuwissen, T. H. E. (2011). Using the genomic relationship matrix to predict the accuracy of genomic selection. *Journal of Animal Breeding and Genetics*, 128, 409–421.

- Gorjanc, G., Bijma, P., & Hickey, J. M. (2015). Reliability of pedigree-based and genomic evaluations in selected populations. *Genetics Selection Evolution*, 47, 65.
- Habier, D., Fernando, R. L., & Garrick, D. J. (2013). Genomic BLUP decoded: A look into the black box of genomic prediction. *Genetics*, 194, 597–607.
- Hayes, B. J., Bowman, P. J., Chamberlain, A. J., & Goddard, M. E. (2009). Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science*, 92, 433–443.
- Hayes, B. J., Visscher, P. M., & Goddard, M. E. (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genetics Research (Cambridge)*, 91, 47–60.
- Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31, 423–447.
- Jannink, J.-L., Lorenz, A. J., & Iwata, H. (2010). Genomic selection in plant breeding: From theory to practice. *Briefings in Functional Genomics*, 9, 166–177.
- Liu, Z., Reinhardt, F., Bünger, A., & Reents, R. (2004). Derivation and calculation of approximate reliabilities and daughter yield-deviations of a random regression test-day model for genetic evaluation of dairy cattle. *Journal of Dairy Science*, 87, 1896–1907.
- Luan, T., Woolliams, J. A., Lien, S., Kent, M., Svendsen, M., & Meuwissen, T. H. E. (2009). The accuracy of Genomic Selection in Norwegian red cattle assessed by cross-validation. *Genetics*, 183, 1119–1126.
- Meuwissen, T., Hayes, B., & Goddard, M. (2013). Accelerating Improvement of Livestock with Genomic Selection. *Annual Review of Animal Biosciences*, 1, 221–237.
- Mrode, R. A., & Swanson, G. J. T. (2004). Calculating cow and daughter yield deviations and partitioning of genetic evaluations under a random regression model. *Livestock Production Science*, 86, 253–260.
- Rincint, R., Laloë, D., Nicolas, S., Altmann, T., Brunel, D., Revilla, P., ... Moreau, L. (2012). Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: Comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). *Genetics*, 192, 715–728.
- Saatchi, M., McClure, M. C., McKay, S. D., Rolf, M. M., Kim, J., Decker, J. E., ... Taylor, J. F. (2011). Accuracies of genomic breeding values in American Angus beef cattle using K-means clustering for cross-validation. *Genetics Selection Evolution*, 43, 40.
- Sargolzaei, M., & Schenkel, F. S. (2009). QMSim: A large-scale genome simulator for livestock. *Bioinformatics*, 25, 680–681.
- Szyda, J., Ptak, E., Komisarek, J., & Żarnecki, A. (2008). Practical application of daughter yield deviations in dairy cattle breeding. *Journal of Applied Genetics*, 49, 183–191.
- VanRaden, P. M. (2007). Genomic Measures of Relationship and Inbreeding. *Interbull Bulletin*, 37, 33–36.
- VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91, 4414–4423.
- Wellmann, R., Preuß, S., Tholen, E., Heinkel, J., Wimmers, K., & Bennewitz, J. (2013). Genomic selection using low density marker panels with application to a sire line in pigs. *Genetics Selection Evolution*, 45, 28.
- Wientjes, Y. C. J., Veerkamp, R. F., & Calus, M. P. L. (2013). The effect of linkage disequilibrium and family relationships on the reliability of genomic prediction. *Genetics*, 193, 621–631.

## SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

**How to cite this article:** Ni G, Kipp S, Simianer H, Erbe M. Accuracy of genomic breeding values revisited: Assessment of two established approaches and a novel one to determine the accuracy in two-step genomic prediction. *J Anim Breed Genet*. 2017;134:242–255. <https://doi.org/10.1111/jbgs.12273>

## APPENDIX 1

### Derivation of the proposed formula to estimate the accuracy of genomic prediction

Let  $t$ ,  $e$  and  $g$  denote the standardized transformed vectors TBV, EBV and GBV for individuals in a population, that is

$$t = \frac{T - \mathbf{1}_n \bar{T}}{\sqrt{\text{var}(T)}}, \quad e = \frac{E - \mathbf{1}_n \bar{E}}{\sqrt{\text{var}(E)}}, \quad \text{and} \quad g = \frac{G - \mathbf{1}_n \bar{G}}{\sqrt{\text{var}(G)}};$$

where  $\bar{T}$  ( $\bar{E}$ ,  $\bar{G}$ ) is the mean of  $T$  ( $E$ ,  $G$ ), respectively.  $n$  is the length of vectors  $T$ ,  $E$  and  $G$ , and  $\mathbf{1}_n$  is a column vector of 1s of length  $n$ . As such a scaling procedure does not affect correlations, the following applies:

$$r_{GT} \equiv r_{gt}, \quad r_{ET} \equiv r_{et}, \quad \text{and} \quad r_{EG} \equiv r_{eg}$$

With  $t$ ,  $e$  and  $g$  being standardized vectors, the sums of their elements are equal to 0, the variances of  $t$ ,  $e$  and  $g$  are equal to 1 and the correlation coefficients are identical to the regression coefficients.

Thus, we can define the following regression:

$$t = b_{t|e}e + m_{et}$$

which can be rewritten as

$$t = r_{et}e + m_{et}$$

with  $b_{t|e}$  being the regression coefficient of  $t$  on  $e$  which equals the correlation coefficient  $r_{et}$  of  $t$  and  $e$ . The vector of standardized TBV  $t$  is therefore expressed as its expectation given the standardized estimated breeding values  $e$  and an error term, denoted as  $m_{et}$ . Note that by definition  $e$  and  $m_{et}$  are uncorrelated, that is,  $\text{cor}(e, m_{et}) = 0$ . In the same way, we can write  $g = r_{eg}e + m_{eg}$  with  $\text{cor}(e, m_{eg}) = 0$ .



Then, the accuracy of GBV can be written as

$$r_{GT} \equiv r_{gt} \equiv \frac{\text{cov}(r_{et}\mathbf{e} + \mathbf{m}_{et}, r_{eg}\mathbf{e} + \mathbf{m}_{eg})}{\sqrt{\text{var}(\mathbf{g})\text{var}(\mathbf{t})}}$$

As  $\text{var}(\mathbf{g}) = \text{var}(\mathbf{t}) = 1$ , it follows

$$r_{GT} \equiv \text{cov}(r_{et}\mathbf{e} + \mathbf{m}_{et}, r_{eg}\mathbf{e} + \mathbf{m}_{eg}) \equiv r_{eg}r_{et}\text{cov}(\mathbf{e}, \mathbf{e}) + r_{et}\text{cov}(\mathbf{e}, \mathbf{m}_{eg}) + r_{eg}\text{cov}(\mathbf{e}, \mathbf{m}_{et}) + \text{cov}(\mathbf{m}_{et}, \mathbf{m}_{eg})$$

Because  $\mathbf{e}$  and  $\mathbf{m}_{et}$  (as well as  $\mathbf{e}$  and  $\mathbf{m}_{eg}$ ) are uncorrelated,  $\text{cov}(\mathbf{e}, \mathbf{m}_{eg}) = \text{cov}(\mathbf{e}, \mathbf{m}_{et}) = 0$ . Thus,

$$r_{GT} \equiv r_{eg}r_{et} + \text{cov}(\mathbf{m}_{et}, \mathbf{m}_{eg}) \equiv r_{eg}r_{et} + r_{m_{et}m_{eg}}\sqrt{\text{var}(\mathbf{m}_{et})\text{var}(\mathbf{m}_{eg})}$$

with  $r_{m_{et}m_{eg}}$  being the correlation between  $\mathbf{m}_{et}$  and  $\mathbf{m}_{eg}$  and being denoted as  $\omega$  in the following.

$\text{var}(\mathbf{m}_{et})$  is

$$\text{var}(\mathbf{m}_{et}) = \text{var}(\mathbf{t} - r_{et}\mathbf{e}) = \text{var}(\mathbf{t}) + r_{et}^2\text{var}(\mathbf{e}) - 2r_{et}\text{cov}(\mathbf{t}, \mathbf{e}),$$

As  $\text{var}(\mathbf{t}) = \text{var}(\mathbf{e}) = 1$ ,

$$\text{var}(\mathbf{t}) + r_{et}^2\text{var}(\mathbf{e}) - 2r_{et}\text{cov}(\mathbf{t}, \mathbf{e}) = 1 + r_{et}^2 - 2r_{et}r_{et} = 1 - r_{et}^2$$

and analogously

$$\text{var}(\mathbf{m}_{eg}) = 1 - r_{eg}^2$$

This results in

$$r_{GT} \equiv r_{eg}r_{et} + \omega\sqrt{(1 - r_{et}^2)(1 - r_{eg}^2)} \equiv r_{EG}r_{ET} + \omega\sqrt{(1 - r_{ET}^2)(1 - r_{EG}^2)} \quad (7)$$