CERES – *Health & Education Medical Journal*
ISSN: 2965-5722
2025, Vol. 3, N°. 2, p. 01-13
DOI:10.62234/ceresv3n2-003

The generalized relative effect (GRE): its introduction and a tutorial in R

El efecto relativo generalizado (ERG): su introducción y un tutorial en R

Jimmie Leppink [1]
[1] Leppink Analytics, Ondara – Alicante, Spain. Orcid: https://orcid.org/0000-0002-8713-1374

Autor Correspondente:
Jimmie Leppink
Calle Hortetes, 5, Piso 2, Puerta 11, 03760 Ondara – Alicante, Spain. E-mail: leppinkanalytics@gmail.com

ABSTRACT

Introduction: When comparing two samples on an outcome of interest, having a statistic that can (1) be applied to all levels of measurements, (2) account for order information in ordinal and quantitative variables, (3) provide valid outcomes for varying sample sizes, and (4) account for prior information from theory or previous research facilitates quantitative and mixed-methods research. Although several non-parametric statistics are available, they all fall short in at least one of these areas. **Objective:** The aim of this paper is twofold: (1) to introduce a non-parametric method called 'generalized relative effect' (GRE) that performs well in all of the aforementioned areas, and (2) to provide a tutorial of how to compute a GRE point estimate and its corresponding credible interval in R. **Theoretical Framework:** GRE unites the strengths of a Bayesian variant of the percentage of all non-overlapping data (PAND-B) and the Brunner-Munzel relative effect (RE). On the one hand, RE can provide valid outcomes for dichotomous, multicategory ordinal, and quantitative outcomes for equal and unequal sample sizes, but cannot be applied to multicategory nominal outcomes and does not account for prior information. On the other hand, PAND-B is applicable to all levels of measurement and accounts for prior information, but does not adequately account for order information in multicategory ordinal and quantitative outcomes, and tends to exaggerate differences when sample sizes are unequal. **Method:** Through a simulated example with a four-category ordinal outcome and clearly unequal sample sizes, this paper compares PAND-B, RE, and GRE. **Results and Discussion:** GRE unites the best of PAND-B and RE and avoids the weaknesses of these statistics. GRE is easy to implement in basic R-code, and guidelines for the interpretation of GRE outcomes in terms of 'small', 'medium', and 'large' are discussed. **Conclusion:** Given its applicability to all levels of measurement and to both equal and unequal sample sizes, GRE provides a consistent approach to the interpretation of the magnitude of differences between two samples in qualitative, quantitative, or mixed sets of outcomes.

HISTÓRIA DO ARTIGO
Received 30 April 2025
Accepted 6 May 2024

KEYWORDS
Statistics; effect size; non-parametric methods; Bayesian methods; levels of measurement

RESUMEN

Introducción: Cuando se comparan dos muestras en un resultado de interés, disponer de una estadística que pueda (1) aplicarse a todos los niveles de medición, (2) tener en cuenta la información de orden en variables ordinales y cuantitativas, (3) proporcionar resultados válidos para muestras de distintos tamaños y (4) tener en cuenta la información previa de la teoría o de investigaciones anteriores facilita la investigación cuantitativa y con métodos mixtos. Aunque existen varias

PALABRAS CLAVE
Estadística; tamaño del efecto; métodos no paramétricos; métodos bayesianos; niveles de medición

estadísticas no paramétricas, todas se quedan cortas en al menos una de estas áreas. **Objetivo:** Este artículo tiene un doble objetivo: (1) introducir un método no paramétrico llamado 'efecto relativo generalizado' (ERG) que funciona bien en todas las áreas mencionadas y (2) proporcionar un tutorial de cómo calcular una estimación puntual del ERG y su correspondiente intervalo de credibilidad en R. **Marco teórico:** El ERG aúna los puntos fuertes de una variante bayesiana del porcentaje de todos los datos no solapados (PAND-B) y el efecto relativo (ER) de Brunner-Munzel. Por un lado, el ER puede proporcionar resultados válidos para resultados dicotómicos, ordinales multicategoría y cuantitativos para tamaños de muestra iguales y desiguales, pero no puede aplicarse a resultados nominales multicategoría y no tiene en cuenta la información previa. Por otro lado, PAND-B es aplicable a todos los niveles de medición y tiene en cuenta la información previa, pero no tiene en cuenta adecuadamente la información de orden en los resultados ordinales multicategoría y cuantitativos, y tiende a exagerar las diferencias cuando los tamaños de las muestras son desiguales. **Método:** A través de un ejemplo simulado con un resultado ordinal de cuatro categorías y tamaños de muestra claramente desiguales, este trabajo compara PAND-B, el ER y el ERG. **Resultados y discusión:** El ERG reúne lo mejor de PAND-B y del ER y evita las debilidades de estas estadísticas. El ERG es fácil de implementar en código R básico, y se presentan directrices para la interpretación de los resultados del ERG en términos de 'pequeño', 'medio' y 'grande'. **Conclusión:** Dada su aplicabilidad a todos los niveles de medición y a tamaños de muestra iguales y desiguales, el ERG proporciona un enfoque coherente para la interpretación de la magnitud de las diferencias entre dos muestras en resultados cualitativos, cuantitativos o mixtos de resultados.

INTRODUCTION

Comparisons of two samples on one or several outcomes of interest – qualitative, quantitative or mixed – constitute an indispensable part of health, medicine, education, and many other fields of research. Although many parametric and non-parametric statistics are available to researchers for such comparisons, they all rely on assumptions and have their strengths and weaknesses. At least four desirable features of statistics for two-sample comparisons can be distinguished.

Key Features of Statistics for Two-Sample Comparisons

To start, outcomes of interest can be of nominal (e.g., blue, white, or red shoes), ordinal (e.g., poor, borderline, satisfactory, or excellent performance), interval (e.g., scores on an intelligence test, assuming that an absolute '0' is not a possible outcome), or ratio (e.g., stock prices) level of measurement.[1] Nearly all statistics for two-sample comparisons have in common that they cannot be applied to outcomes of at least one of these levels of measurement.

A second desirable feature of a statistic for two-sample comparisons, at least as far as multicategory ordinal and quantitative (i.e., interval or ratio level of measurement) variables are considered, is an appropriate treatment of the order information. For example, excellent performance is superior to satisfactory performance, satisfactory performance is better than borderline performance, and the latter is still somewhat (although perhaps not much) better than poor performance; a statistic that is supposed to be useful for multicategory ordinal (and/or quantitative) outcomes should account for such order information. On the contrary, a statistic that treats the aforementioned performance

outcome as nominal pretends that the four performance categories are distinguishable categories that cannot be ordered in terms of least to best, and this inappropriate treatment of an ordinal outcome tends to come at cost of incorrect conclusions. The only exception can be found in dichotomous outcomes; given a transparent and consistent coding (e.g., 0/1 dummy coding), the order of categories – at least for the analysis in question – is clear, and differences between two samples in the dichotomous outcome should be interpretable in an unequivocal manner.

In addition the level of measurement and order information features, a statistic for two-sample comparisons should provide valid outcomes regardless of whether the samples are of equal or unequal size. While many of the statistics currently available for two-sample comparisons meet this requirement, as is discussed in the next section, not all statistics encountered in the literature do.

Finally, since research is usually about contributing to theory and research available, and is in the process commonly influenced *by* theory and previous research, the ability to incorporate information from theory and/or previous research in the analysis of group differences is a desirable feature of any statistic.[2]

Key Statistics for Two-Sample Comparisons

Although researchers have a variety of statistics for two-sample comparisons at their disposal, they all fall short in one or several of the previously discussed four desirable features. However, some of the statistics available are promising, especially when they are combined with specific alternatives.

On the one hand, Karch [3,4] provides convincing arguments for why Brunner-Munzel's test (BM) should be used instead of the widely used Mann-Whitney test, because several assumptions underlying the latter are not realistic and their violation can easily result in incorrect conclusions. Key to BM is the relative effect (RE). In a nutshell, every observation in sample A is compared to every observation in sample B, and for each comparison, the outcome is one of the following: higher in B, higher in A, or equal (i.e., a tie or, in sports such as soccer, a 'draw'). For example, if two samples of $n = 10$ each are compared, there are 10 x 10 = 100 comparisons. Suppose, 60 of these comparisons are in favor of sample A (i.e., higher in A), 30 are in favor of sample B (i.e., higher in B), and the remaining 10 comparisons result in a tie. The ties are equally divided over the two samples, resulting in 65 comparisons in favor of A and 35 comparisons in favor of B. For this comparison, RE equals 65 / 100 = 0.65 if formulated in favor of A and 35 / 100 = 0.35 if formulated in favor of B. Either way, there is a 65% chance that a randomly selected observation in sample A is higher than a randomly selected observation in sample B, and a 35% chance that a randomly selected observation in sample B is higher than a randomly selected observation in sample A. As such, RE provides a fairly easily interpretable effect size statistic for two-sample comparisons on dichotomous, multicategory ordinal, and quantitative outcomes. However, since multiple nominal categories cannot be ordered in a non-arbitrary manner, BM and its RE cannot be applied to multicategory nominal outcomes, unless that multicategory nominal outcome is converted into 0/1 dummies for each category and RE is calculated for each category, which would pose challenges to the interpretation of outcomes. Finally, the incorporation of prior information is not currently an option in BM and its RE.

On the other hand, a commonly used statistic in some areas of research is the percentage of all non-overlapping data (PAND).[5] This statistic answers the question what degree of non-overlap there is between two samples on an outcome of interest. For example, if in a group of 20 students (sample A) 19 wear blue shoes and one wears red shoes, and in a second group of 20 students (sample B) 19 wear red shoes and one wears blue shoes, we are only two data points away from perfect non-overlap:

the student with red shoes in sample A and the student with blue shoes in sample B. Consequently, PAND is 95% or 0.95. Given a transparent and consistent coding, it can be demonstrated that in the case of dichotomous outcomes and equal sample sizes, RE and PAND yield the same point estimate and that Pearson's and Spearman's correlation coefficient are exactly or approximately equal to: (2 x PAND) – 1. In other words, correlations of -0.5, 0, 0.1, 0.3, and 0.5 then correspond – at least approximately – with PAND and RE values of 0.25, 0.5, 0.55, 0.65, and 0.75 respectively.

However, it can also easily be demonstrated that when sample sizes are unequal, the correlation coefficient is still fairly close to '(2 x RE) – 1', yet PAND exaggerates the difference. Consider the following example: in class A, 21 students fail and 7 students pass the exam, whereas in class B, 9 students fail and 3 students pass the exam. For this example, we find a correlation of 0 and RE = 0.5, which is correct since the pass/fail distribution is 1:3 in both classes. However, PAND equals 0.6 in this case, as if we found RE = 0.6 and a correlation of about 0.2 (!). Although a Bayesian version of PAND – coined PAND-B – which treats PAND as a binomial variable (i.e., each observation can be a non-overlapping 'success' or an overlapping 'failure' data point) and adds a Binomial prior distribution, helps to account for prior information (in the absence of prior knowledge: one success and one failure) [6,7], it does not circumvent the problem of PAND in the face of unequal sample sizes. In addition, although it can provide a useful effect size statistic for all levels of measurement including nominal as long as sample sizes are equal [7], it does not account for order information as RE does. In other words, although in the case of equal sample sizes PAND and RE provide the same point estimate when outcomes are dichotomous, RE can be expected to perform more adequately when two samples are compared on multicategory ordinal or quantitative outcomes.

In sum, although advantages of PAND-B over RE are the possibility to incorporate prior information and its applicability to outcomes of all levels of measurement including multicategory nominal outcomes, RE is to be preferred when sample sizes are unequal and/or when outcomes are multicategory ordinal or quantitative. If we agree on the four desirable features of a statistic for two-sample comparison presented at the start of this article, a question that arises is if there is a way to unite the relative strengths – and avoid the limitations – of PAND-B and RE in a single statistic. And this is where the generalized relative effect (GRE) is introduced.

Uniting the Strengths and Avoiding the Limitations of Statistics Currently Available

Given the equal division of ties to both samples in the computation of RE, all pairwise comparisons of observations are essentially reduced to being in favor of either one sample or the other. However, the number of pairwise comparisons is much larger than the total sample size; it equals product of the two sample sizes. As in an earlier example, when two samples have $n = 10$, there are 10 x 10 = 100 comparisons. Alternatively, if one sample has $n = 5$ and the other sample $n = 20$, we also have 100 comparison, albeit via a different route: 5 x 20. Consequently, to rescale the numbers in favor of sample A and in favor of sample B (and ties, which are ultimately equally divided over A and B) to the original sample size, we have to divide the counts resulting from all pairwise comparisons by the ratio of the product of the sample sizes (e.g., 100 in the case of two samples of size 10) and the sample size itself (in this case: 20). In the earlier example of 65 comparisons in favor of A and 35 comparisons in favor of B, that means dividing 65 and 35 by (100 / 20), resulting in 13 and 7, respectively. These numbers correspond with proportions of 0.65 and 0.35, yet on a scale of a sample size of 20 instead of on a count of 100 comparisons. Adding a Binomial prior distribution (in the absence of prior knowledge: one success and one failure) analogous to PAND-B [6], we have a Bayesian alternative to

BM's RE with a point estimate close to that of RE (the prior has some influence, though less with increasing sample sizes) and a 90% or 95% credible interval similar to the 90% or 95% confidence interval around the RE point estimate. In this case, if using a default '1 success (A > B) and 1 failure (B > A)' prior distribution, the resulting posterior distribution would be:

B(13,7) + B(1,1) = B(14,8).

This is a posterior distribution with a median (point estimate) of 0.641 and a 95% credible interval ranging from 0.430 to 0.819, which is a wide interval indeed, since the sample size in this example is only $N = 20$. The interval includes 0.5, which represents the default null hypothesis (H_0) that the groups are comparable (i.e., 50% chance of 'B > A' and 50% chance of 'A > B', like in a soccer match where, given the data of recent performance at hand, both teams have the same chance of winning).

In addition, the point estimate is not exactly 0.65, because the non-informative prior of B(1,1) has a median of 0.5 and hence pulls the point estimate slightly towards 0.5. However, as explained earlier, the weight of the prior relative to the data decreases with increasing sample size, as also becomes clear in an example later in this article.

Just like for RE, the previously described computational process for GRE works for multicategory ordinal, quantitative, and – provided a transparent and consistent coding – dichotomous outcomes. However, since for multicategory nominal variables no single unarbitrary coding system – other than, as discussed previously, creating dummy variables for each category and computing for each dummy variable – can be established, GRE must be computed as follows. Suppose, the 60 'A > B', 30 'B > A', and 10 'A = B' example introduced earlier was for a multicategory nominal outcome: we would in that case not be able to distinguish which differences mean 'A > B' and which ought to be understood as 'B > A'. In this case, instead of equally dividing the ties over 'B > A' and 'A > B', all comparisons resulting in 'A ≠ B' are to be interpreted as A and B being *different*, whereas all comparisons resulting in 'A = B' are coded as A and B being the *same*. For the example at hand, this would mean 90 times 'different' and 10 times 'same', which rescaled back to the original sample size (i.e., dividing by 100 / 20) gives us 18 and 2, respectively. In this case, if using a default '1 success (A ≠ B) and 1 failure (B = A)' prior distribution, the resulting posterior distribution would be:

B(18,2) + B(1,1) = B(19,3).

This is a posterior distribution with a median (point estimate) of 0.875 and a 95% credible interval ranging from 0.696 to 0.970. However, note that the interpretation of this point estimate and interval is different from the previous one: in this case, it is not about the chance of a random observation in one sample exceeding a random observation from the other sample but about the chance of a random observation in one sample being *different* from a random observation from the other sample. Especially for quantitative outcomes for which a range of values are possible, high levels of difference may be expected. In addition, this second interval does not take into account the order information in multicategory ordinal and quantitative outcomes. While especially when an ordinal outcome has only a limited number of categories (e.g., three or four), this second interval can help to provide useful information regarding the degree of *ties* (i.e., same) in the set of pairwise comparisons, the main interest when dealing with an ordinal outcome (as for a quantitative outcome and consistently coded

dichotomous outcome) usually lies 'better vs. worse' interpretations rather than just any differences regardless of their directions.

In the 'Methods and Results' section of this article, both computational processes are explained with a simulated data example and basic R-code that can be used to acquire posterior medians and X% (e.g., 95%, 90% or 89%) credible intervals for each type of outcome. Moreover, to compare PAND-B, RE, and GRE on the four desired features of a statistic for two-sample comparisons outlined at the start of this article, the example used in the remainder of this article is one of a comparison of two sample of clearly unequal size on a four-category ordinal outcome.

METHODS AND RESULTS

Suppose, a sample of 130 students is randomly divided into a control condition and an experimental treatment condition. In the control condition, students prepare for an OSCE station the way they usually do, whereas students in the experimental treatment condition are prompted to use an innovative reasoning approach. After training (i.e., the preparation), all students complete the same OSCE station and are rated by an independent rater (who does not know about the two conditions or which student was in which condition) among others on the following overall performance scale: *poor*, *borderline*, *satisfactory*, or *excellent* performance. Although this kind of scales is frequently treated as if it was of interval level of measurement, it is actually an ordinal outcome. After all, what guarantee do we have that the difference between each two subsequent categories of performance is always the same? In addition, for the study at hand, the research only have resources for 30 students in the experimental treatment condition, hence the random allocation to conditions is done such that 100 students end in the control condition. Suppose, the outcomes are as presented in Table 1.

Table 1. Outcomes of the hypothetical example study.

		Control	**Experimental**
		***n* = 100 (76.9%)**	***n* = 30 (23.1%)**
Performance	Poor	5 (5.0%)	1 (3.3%)
	Borderline	20 (20.0%)	4 (13.3%)
	Satisfactory	65 (65.0%)	18 (60.0%)
	Excellent	10 (10.0%)	7 (23.3%)

The cell percentages indicate proportionally slightly fewer 'poor' and 'borderline' ratings and somewhat more 'excellent' ratings in the experimental treatment condition. Spearman's rho and Kendall's Tau-B coefficient are 0.149 and 0.142, respectively, positive because there is a slight tendency towards differences in favor of the experimental treatment condition rather than the other way around.

For RE, thanks to Karch[4], there is user-friendly package in *jamovi*[8], which for the data at hand provides a point estimate of 0.588, a 95% confidence interval of [0.480; 0.695], and a 90% confidence interval of [0.498; 0.677]. Applying the previously introduced '(2 x PAND) – 1' rule of thumb, this would translate to a correlation of 0.176, which coincides with neither Spearman's nor Kendall's coefficient but is close enough and would result in the same kind of interpretation in terms of the strength of the association: rather weak.

Since PAND and its Bayesian counterpart PAND-B are about the easiest way to achieving perfect non-overlap[5-7] and in the case of dichotomous, multicategory ordinal or quantitative outcomes in a given direction (here: in favor of the experimental treatment condition),[6] eliminating or swapping the 'poor', 'borderline', and 'satisfactory' observations in the experimental treatment group (i.e., 1 +

4 + 18) and the ‘excellent’ observations in the control group (i.e., 10) would result in perfect non-overlap. In other words, there are 1 + 4 + 18 + 10 = 33 observations that are contributing to overlap, hence PAND equals:

PAND = (130 – 33) / 130 = 0.746.

Applying the previously introduced ‘(2 x PAND) – 1’ rule of thumb, this would translate to a correlation of 0.492, which of course for the data at hand is way off. And for PAND-B, we would find:

B(97,33) + B(1,1) = B(98,34).

This is a posterior distribution with a median (point estimate) of 0.744 and a 95% credible interval ranging from 0.665 to 0.813. Applying the previously introduced ‘(2 x PAND) – 1’ rule of thumb, this would translate to a correlation of 0.488, which is still far away from where a correlation computed on this dataset should be.

However, for GRE, we proceed as follows. First, in **Step 1**, we have to read a data file, for example in ‘.csv’-format, and define our variables:

```
# data
data <- read.csv("file-location")
group <- data$group
value <- data$value
```

Next, in **Step 2**, we have to split our data set into two groups and determine the total sample size (here $N = 130$) in order to determine the number of pairwise comparisons (here: 100 x 30 = 3,000) and subsequently rescale back to the original sample size as explained previously:

```
# split data
c <- subset(data, group == "c")
e <- subset(data, group == "e")
# define sample size
k <- nrow(data)
```

At this point, in **Step 3**, we run all pairwise comparisons, calculate a difference statistic, and categorize each difference in terms of ‘Experimental > Control’ (i.e., positive, in the following code ‘pos’), ‘Control > Experimental’ (i.e., negative, in the following code ‘neg’), or tie:

```
# all pairwise combinations
pairwise <- merge(c, e, by = NULL, suffixes = c("_c", "_e"))
# pairwise difference
difference <- pairwise$value_e - pairwise$value_c
# three categories
pos <- ifelse(difference > 0, 1, 0)
```

```
tie <- ifelse(difference == 0, 1, 0)
neg <- ifelse(difference < 0, 1, 0)
```

Next, in **Step 4**, we need to define our prior distribution, B(1,1) in the absence of any prior knowledge:

```
# prior parameters
a <- 1
b <- 1
```

Penultimate, in **Step 5**, we equally divide the ties over 'positive' (in favor of experimental) and 'negative' (in favor of control) differences and rescale back to the original sample size:

```
# assuming ordinal categories, in favor of experimental (e)
# pairwise comparison data scaled back to the original sample size
x <- k * (mean(pos) + 0.5 * mean(tie))
n <- k
re <- x / n
```

And finally, in **Step 6**, we obtain the posterior distribution with its median (i.e., the GRE point estimate) and desired X% credible interval (although the below example demonstrates the 95%, any other desired X% credible interval can be computed as well):

```
# posterior parameters
post_a <- a + x
post_b <- b + n - x
# posterior median
posterior_median <- qbeta(0.5, post_a, post_b)
posterior_median
# the 95% credible interval
e_ci95 <- qbeta(c(0.025, 0.975), post_a, post_b)
e_ci95
```

Doing so, we find a GRE point estimate (i.e., posterior median) of 0.587, which nearly coincides with the RE point estimate. Applying the previously introduced '(2 x PAND) – 1' rule of thumb, this would translate to a correlation of 0.174, which coincides with neither Spearman's (0.149) nor Kendall's (0.142) coefficient but is close enough and would result in the same kind of interpretation in terms of the strength of the association: rather weak. Further, as for the interval estimation, we find a 95% credible interval of [0.501; 0.668], a 90% credible interval of [0.515; 0.656], and an 89% credible interval of [0.517; 0.654]. Due to the prior distribution, an X% credible interval is slightly narrower than its Frequentist X% confidence interval counterpart.

Finally, if these four performance categories were not ordinal but nominal, or in addition to a 'better-worse' tendency we are interested in the distribution of 'different' vs. 'same' ratings, only the *first line of code* in **Step 5** is *slightly* different:

```
# pairwise comparison data scaled back to the original sample size
x <- k * (mean(pos) + mean(neg))
n <- k
re <- x / n
```

That is, instead of 'x <- k * (mean(pos) + 0.5 * mean(tie))', we write 'x <- k * (mean(pos) + mean(neg))', because in this second variant we are interested in the proportion of *differences* rather than in a 'better-worse' tendency. With this minor difference in one line of coding – nothing more than replacing '0.5 * mean(tie))' by 'mean(neg))', we obtain a GRE point estimate and the desired X% credible interval for different vs. same, which is a variant of GRE that can be used for *all types of outcomes* including multicategory nominal ones. For the example at hand, this results in a posterior median (i.e., GRE point estimate) of 0.558, a 95% credible interval of [0.472; 0.641], a 90% credible interval of [0.486; 0.628], and an 89% credible interval of [0.488; 0.626].

To conclude, although from an *ordinal* perspective, the 95% credible interval for GRE is entirely above H_0: $p = 0.5$ (i.e., the two groups are comparable in performance), indicating a better performance in the experimental treatment condition, if from a *nominal* perspective we were to test H_0: $p = 0.5$ (i.e., 50% chance of a difference, 50% chance of a tie), we would conclude insufficient evidence that there is more than a 50% chance of a difference.

DISCUSSION

This article started from the arguments that (1) two-sample comparisons are very common in many fields of research in qualitative, quantitative, and mixed-methods studies, and (2) a statistic for two-sample comparisons should be applicable to all levels of measurement, account for order information in ordinal and quantitative outcomes, provide valid outcomes for varying sample sizes, and account for prior information. To date, RE provides valid results for dichotomous, multicategory ordinal, quantitative outcomes for equal and unequal sample sizes but is not applicable to multicategory nominal outcomes in a straightforward manner. And although PAND and PAND-B are applicable to the latter and PAND-B allows for the incorporation of prior information, both non-overlap statistics fail to adequately account for order information and clearly fall short in the face of unequal sample sizes. GRE unites the strengths of RE and PAND-B and circumvents the limitations of both.

Although GRE in essence follows the computational process of BM's RE (i.e., a pairwise comparison of each observation in one sample with each observation in the other sample), it adds a prior distribution and, with only a very minor tweak (i.e., replacing one term in one line of code), it can be used for multicategory nominal outcomes as well. Given its proximity to RE, GRE accounts for order information conform RE and in the face of unequal sample sizes does not produce strange estimates such as the ones returned by PAND and PAND-B. Applying the previously introduced '(2 x PAND) – 1' rule of thumb, GRE estimates of 0.25, 0.35, 0.45, 0.50, 0.55, 0.65, and 0.75 approximately correspond with correlations of -0.5, -0.3, -0.1, 0, 0.1, 0.3, and 0.5, respectively. Hence, if in a given research context, we interpret correlations around 0.1, 0.3, and 0.5 as 'small', 'medium', and 'large', respectively, GRE values around 0.55 indicate 'small' differences, GRE values around 0.65 are indicative of 'medium' differences, and GRE values around 0.75 can be interpreted as relatively 'large' differences. Moreover, in addition to the point estimates, credible intervals can be used to indicate ranges of plausible hypotheses.

Contrary to *t*-tests and other parametric approaches, RE and GRE can account for order information without requirements of Normal (i.e., bell-shaped) and homoscedastic (i.e., equal variance) distributions. That said, when applied to quantitative outcomes, the distribution of pairwise comparison differences can be visualized (e.g., histograms, boxplots, or clouds) to inspect the degree of asymmetry in that distribution. If two approximately normally distributed samples are compared, the distribution of pairwise comparison differences can be expected to approach a Normal distribution as well, whereas strong skewness and/or extreme cases in at least one of the samples will usually be reflected in one or both tails of the distribution of pairwise comparison differences as well.

Furthermore, although the example in this article uses a one-way design – and further simulations and studies including real data could be dedicated to questions like performance and statistical power of GRE under different circumstances – GRE can be applied to balanced 2 x 2 designs as well. After all, in a balanced 2 x 2 design, all four cells have equal sample sizes, and the main effect of each factor as well as their interaction effect can be estimated independently. The main effect of factor A is then computed by comparing the two cells where A = 1 vs. the two cells where A = 0, the main effect of factor B is computed by comparing the two cells where B = 1 vs. the two cells where B = 0, and one of several ways to compute the interaction effect is to contrast '[A = 0 and B = 0] and [A = 1 and B = 1]' with '[A = 0 and B = 1] and [A = 1 and B = 0]'. This way, GRE can – at least for balanced 2 x 2 designs – provide point and interval estimates for both main effects and the A-by-B interaction effect regardless of whether the outcome is nominal, ordinal, or quantitative. This generalization does not hold for clearly unbalanced designs, because in the latter there is design correlation between the factors, and hence main and interaction effects can no longer be estimated independently.

Finally, GRE can also provide a promising approach to multigroup comparisons in one-way designs, either by applying GRE to all possible pairs of groups (here, many statisticians will argue for some kind of correction for multiple testing) or – if a specific directed alternative hypothesis is at hand – to specific group comparisons. This extension as well as that for balanced 2 x 2 designs should be subjected to further research, among others because comparisons of multiple groups could create computational challenges, especially in the case of larger samples.

CONCLUSION

Given its applicability to all levels of measurement and to both equal and unequal sample sizes, GRE provides a consistent approach to the interpretation of the magnitude of differences between two samples in qualitative, quantitative, or mixed sets of outcomes. Whether we deal with qualitative findings, quantitative findings, or the interest lies in a mixed-methods integration of qualitative and quantitative findings, prior information can be incorporated, and differences can be interpreted in terms of 'small', 'medium', and 'large'. In addition, its generalization to multigroup comparisons as well as to main and interaction effects in balanced 2 x 2 designs could be investigated further.

ACKNOWLEDGEMENTS

The author wishes to thank Dr. Patricia Pérez-Fuster for insightful conversations about statistics for small samples. These conversations contributed to the development of PAND-B years ago, and laid the foundation for the RE and PAND-B integration into GRE as outlined in this article.

FUNDING

This article and research contributing to it did not receive any funding at any point in time.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

REFERENCES

1. Stevens SS. On the theory of scales of measurement. Sci. 1946;26:677-680. http://www.jstor.org/stable/1671815

2. Lindley D. Bayesian statistics: A review. London: SIAM; 1972.

3. Karch JD. Psychologists should use Brunner-Munzel's instead of Mann-Whitney's U test as the default nonparametric procedure. Adv. Methods Pract. Psychol. Sci. 2021;4(2). https://doi.org/10.1177/2515245921999602

4. Karch JD. bmtest: A Jamovi Module for Brunner–Munzel's Test — A Robust Alternative to Wilcoxon–Mann–Whitney's Test. Psych. 2023;5:386-95. https://doi.org/10.3390/psych5020026

5. Parker RI, Hagan-Burke KJ, Vannest KJ. Percentage of all non-overlapping data (PAND): An alternative to PND. J Spec Educ. 2007;40(4):194-204. https://doi.org/10.1177/00224669070400040101

6. Leppink J. The art of modelling the learning process: Uniting educational research and practice. Cham: Springer; 2020. https://doi.org/10.1007/978-3-030-43082-5

7. Leppink J. Un modelo bayesiano para datos cualitativos en simulación. Rev Latam Sim Clin. 2021;3(3):117-9. https://doi.org/10.35366/103188

8. The jamovi Project. Jamovi (Version 2.6) [Computer Software]; 2024. Retrieved from https://www.jamovi.org