Some Adventures in Reliability Estimation

Anne Thissen-Roe

Abstract In many measurement settings, the reliability of a measure can be straightforwardly estimated. One might have a test that is at least supposed to be unidimensional, on which everyone is supposed to respond to all the same items, and the score is a simple sum of correct answers. Or maybe one has a computerized adaptive test, producing score estimates with individual errors of measurement and corresponding individual reliability values, all based on the same posterior distribution. But sometimes, such as in industry applications of gamified assessment, one arrives at measures for which one wants to estimate the reliability—and yet, they don't look much like traditional tests, or use IRT for scoring. Then, one is obliged to be adventurous. This collection of anecdotes from a career-matching platform features applications of a variety of techniques for estimating reliability in unusual situations, such as composite reliability, structured applications of split-half, and modeling and simulation.

Keywords Reliability · Game-based assessment

1 Background

At pymetrics, we do career matching. We've turned a set of classic psychology and neuroscience experiments into web and phone games. From the gameplay we get a highly multidimensional set of measures, on which we use machine learning to map users' score patterns to job families where people like them are successful. A benefit of this approach is that we can match applicants to careers based on qualities of people from every walk of life. Most measures do not have "good" and "bad" directions, but distinguish the needs of one kind of job task from another,

A. Thissen-Roe (✉)
Pymetrics, Inc., New York, NY, USA
e-mail: anne@pymetrics.com; https://www.pymetrics.ai

M. Wiberg et al. (eds.), *Quantitative Psychology*, Springer Proceedings
in Mathematics & Statistics 393, https://doi.org/10.1007/978-3-031-04572-1_1

so that every form of responding can have a match to some kind of employment. Multidimensional measurement also allows us to optimize for prediction subject to fairness constraints (Wilson et al., 2021). Finally, the two-step approach lets pymetrics do selection, internal mobility, outplacement, a job marketplace, all based on one set of measures.

Thus, we have game-based psychological measurement with substantial plurality of purpose. To support this, the measures need to be robust, reliable, and as broadly as possible valid. The machine learning algorithm for selection models works to ensure that irrelevant or noisy measures aren't used to select people into a job, but even machine learning benefits from good "feature engineering" of its inputs. That is, using the tools of psychometrics, we audit measurement quality and find ways to improve it.

One of the metrics of quality we care about is reliability, or standard error of measurement, or signal to noise ratio. Those are all algebraic transformations of each other, and can be conditional or general depending on the circumstances. Internally at pymetrics, we talk about all three; this chapter discusses the issue primarily in terms of reliability.

Reliability is corroborating evidence for our validity argument. The preferred, primary evidence is a criterion validation study, but reliability helps us be sure that study is not capitalizing on chance. For our client contacts, many of whom have background in industrial-organizational psychology, reliability is a familiar concept and a reassuring metric to have available. In addition, reliability estimates help us evaluate proposed changes to measures, and help us make sense of the factor model we use for our explanations and descriptions. We have measures that legitimately measure non-redundant constructs, but we have also had some plain noisy ones, and reliability information helps us not to confuse the two. Notably, these measurements are delivered to machine learning models and not directly to humans, alleviating the need to communicate uncertainty in an individual score to an end user.

Games being games, our measure set is laden with eccentric scoring algorithms. To estimate reliability, we can't always rely on Cronbach's alpha, or get conditional standard errors from our latent trait estimation procedure. Some of our measures are strange beasts, and estimating their reliability is an adventure.

So this is a collection of short stories of reliability estimation in our applied context. I hope they prove encouraging or useful to somebody else who needs to be creative, or at the very least, they're entertaining.

These anecdotes include a case where we had interdependencies between the value of one measure and the reliability of another, and a case where we had an unusual scoring algorithm combined with extensively counterbalanced stimuli. There are two cases of composite scores, and finally, one case where the best answer came from process modeling.

2 Interdependencies: The Story of Easy or Hard

In modern test scoring, a measure’s reliability may not be a uniform property applicable to all possible persons, scores and response patterns. We are accustomed to the standard error of measurement being higher at extreme scores, or perhaps for atypical response patterns. We can find the conditional standard error of measurement for a user, and then convert it to an individual reliability value applicable to a score; we can profile the quality of measurement across a population or across the range of measurement.

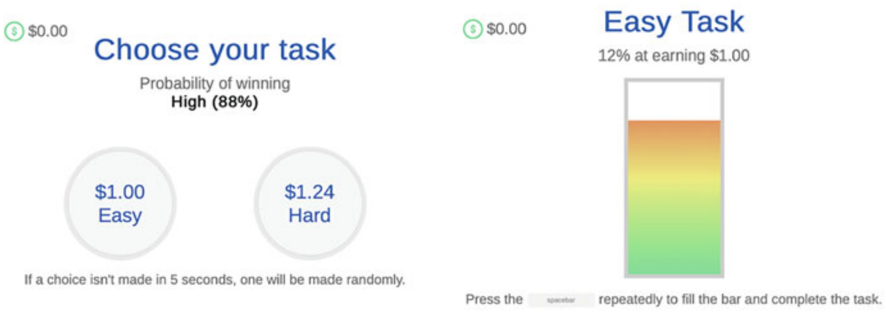
The situation becomes more complicated when the reliability of one measure depends in part on performance on *another* measure. This was the challenge presented by Easy or Hard.

Easy or Hard is adapted for brevity from the Effort-Expenditure for Rewards Task (EEfRT) (Treadway et al., 2009), a popular measure of motivation for rewards developed initially to study effort-based decision-making (Fig. 1). A pymetrics user is presented with a series of choices between an easy task and a hard task. Both tasks involve tapping a key or touch-screen hotspot a certain number of times within a certain interval, but the hard one is more demanding: sixty taps in twelve seconds instead of five times in three seconds. Each choice offers different game money rewards for the two tasks, with an equal probability of payout upon success, regardless of the task chosen. The probability of payout varies across trials. Easy or Hard has nine different expected value differences based on probability and amount of payout.

There is a total time limit of two minutes, after which the game ends. This makes Easy or Hard, in essence, a resource allocation task: a finite quantity of available time must be divided among task-choosing decision time, time to complete easy tasks, and time to complete hard tasks. Hard tasks have higher rewards, but can take up to twelve seconds to complete, and in addition to the random probability of no payout, it is realistically possible to fail the hard tap task. By contrast, it’s essentially impossible to fail the easy task, but it is easy to complete it quickly and move on. The user is not required to wait out the three seconds. Two or three easy task payouts can easily exceed one hard task payout. Further, any time spent deciding between the tasks and their offered rewards is time without pay.

Some users scan the information and then pick a task based on calculation or heuristics. Some users always choose the easy task, and others always choose hard. The decision itself is a speeded task, as well. After five seconds, a task is assigned at random, which is undesirable as a strategy for obtaining rewards.

In addition to neuroscience findings linking hard task selection on the EEfRT to reward processing brain activity (Treadway et al., 2012), the EEfRT fits in a class of measures of extrinsic motivation (Ryan & Deci, 2000). In the workplace, a measure of extrinsic motivation can provide insight into how compensation may motivate employees, in particular when indicators of intrinsic motivation are also considered (Cerasoli et al., 2014). A multidimensional view of Easy or Hard begins to address this construct space.



If a choice isn't made in 5 seconds, one will be made randomly.

Press the spacebar repeatedly to fill the bar and complete the task.

Fig. 1 *Easy or Hard. Left side: Task selection.* The user has five seconds to choose between an "Easy" and a "Hard" task. Either task, if completed successfully, gives the user a specified chance (here, 88%) of receiving a game-money reward. The reward is always larger for the "Hard" task. *Right side: Task completion.* A user selecting the "Easy" task must press the spacebar five times in three seconds in order to have the specified probability (here, 12%) of obtaining a game-money reward. A user selecting the "Hard" task must press the spacebar sixty times in twelve seconds in order to have the same chance of a higher reward

We are therefore, first, interested in a user's probability of selecting the hard task under different expected value conditions, and overall. Some users show a consistent preference for the hard task even when the rewards are low and its selection is not strategic; this may be an indicator of one kind of intrinsic motivation. Another measure is the time taken to choose a task, an indicator of attention paid to the differential reward information.

The challenge in calculating the reliability of either the hard task preference measure or the response time measure is that both are conditional on the number of decision responses contributing to the measure, which is in turn conditional on the number of hard tasks selected (or assigned following timeout) and also the amount of time spent on the decision screen. If a user times out or nearly so, and always chooses or is assigned hard, that user can run out of total time in five tasks. We observe some cases of that in our data. We also see the occasional user that can race through thirty, fifty, or a hundred easy tasks, without reading anything. The content of a user's choices determines the quality of that user's measurement.

It was, however, possible to calculate Cronbach's alpha for the subset of users that reached each item from the first to the 30th, which included fewer users for each additional item after five. We produced the whole table, but when we need a single number for reporting, we prefer the reliability for a user at the median number of trials.

As shown in Fig. 2, the quality of the RT measures improves rapidly in the first few trials, and more slowly thereafter for a long time. Averaging log (RT) is unsurprisingly better than averaging RT, but the same general behavior can be seen in both.

For the probability of choosing hard, however, the reliability dips for additional responses between about ten and fifteen. Around ten responses, the number of

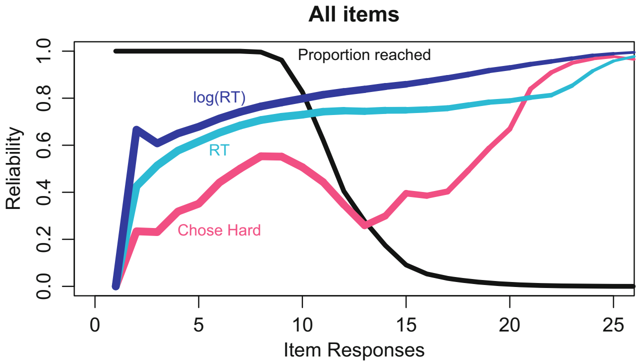


Fig. 2 *Easy or Hard: Reliability as a function of items reached.* Generally, measures of response time (averaged directly or as logarithms) increased in reliability with each additional task completed by the user. However, the reliability of preference for the "Hard" task ("Chose Hard") dipped in reliability between ten and fifteen items reached. This is a range restriction phenomenon; due to the time limit, users can only reach items above the low teens by sometimes selecting "Easy" tasks. The higher the item count, the more "Easy" tasks are generally necessary to reach it. The range of possible scores on the preference measure is limited, and the variance reduced, on the right side of the plot. This in turn suppresses the reliability

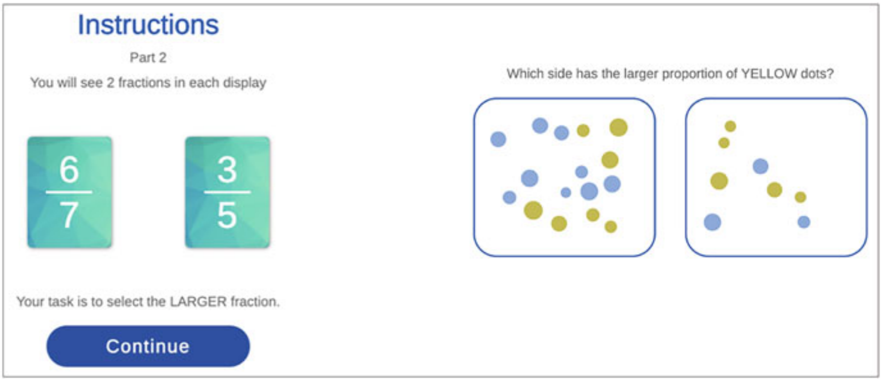
users reaching each trial declines sharply, especially among the users with a high probability of picking hard. As a result, the variance of preference for hard across users drops, and by fifteen items, the complete-data sample is left with only users who chose several easy tasks. Even if the absolute error of measurement stays the same or improves, the reliability drops, relative to the surviving user population at each item. It's a good reminder that reliability isn't solely a property of the test, but of the composition of the sample.

3 Counterbalancing: The Story of Magnitudes

In working with tests that have subject matter constraints, curriculum coverage requirements, and other intentional non-uniformities in item content, there is a need to assess reliability in a way that recognizes that the mixture of content is intentional and would be consistent across hypothetical alternate forms. We frequently see, for example, a stratified version of Cronbach's alpha in use.

What happens, though, when an intentional mixture of items meets a scoring method that isn't a sum score, mean score, or latent trait score? This was the challenge we faced for two of the measures in Magnitudes.

Magnitudes is a pair of measures of Approximate Number Sense (ANS): Dots and Fractions (Fig. 3). Each measure is made up of quantitative comparison items. In Dots, the user selects the side that has the higher proportion of yellow dots. In Fractions, the user selects the larger fraction. They measure nearly the same



Instructions

Part 2

You will see 2 fractions in each display

Your task is to select the LARGER fraction.

Which side has the larger proportion of YELLOW dots?

Fig. 3 *Magnitudes. Left side: Fractions.* A user is presented with two fractions side by side, and must choose the larger. The ratio of the two fractions is manipulated in order to increase or decrease difficulty. *Right side: Dots.* A user is presented two side-by-side arrays of mixed yellow and blue dots, and must select the array with proportionally more yellow dots. The numbers of blue and yellow dots are manipulated so as to contrive a range of ratios of proportions yellow. Automatic item generation is used to produce item clones with the specified dot counts, with the variably-sized dots distributed in a unique, non-overlapping pattern for each user

theoretical construct, but Fractions uses symbolic numbers and Dots uses non-symbolic quantity.

The primary metric of interest is the Weber fraction w for numerosity, which comes from psychophysics, and is the minimum difference between two quantities at which the greater quantity can be reliably recognized by an individual, expressed as a fraction of the smaller quantity. The Weber fraction w of an individual relates to the ratio r of the quantities compared, the improper fraction of the larger over the smaller quantity, which yields 75% accurate performance by that individual (Hunt, 2007):

$$w = r - 1\ . \tag{1}$$

There is a model of performance implied by each of two theories of neural development; we currently use the Linear Spacing Model:

$$P(correct) = \boldsymbol{\Phi}\left(\frac{1}{w} * \frac{(r-1)}{\sqrt{r^2+1}}\right) \tag{2}$$

(Dietrich et al., 2016; Dehaene, 2007).

A separate w score is calculated for each section, because they measure different attributes (Baker & Thissen-Roe, 2021). Although the Weber fraction can be estimated as a latent trait (Thissen-Roe & Baker, 2021), we originally implemented it with estimation using a least squares algorithm (Price et al., 2012). We need to estimate the reliability of the least squares w as a baseline for evaluating the latent trait version, as well as for the validity and documentation purposes common to all pymetrics measures.

For the least squares method, trials were grouped into four difficulty levels, ten trials each. The trials were also carefully counterbalanced for which side, left or right, has the correct response, and to foil heuristics applicable to each section, such as pixels of each color or fractions over one half. This counterbalancing creates an intentional mixture structure, similar to requirements for curriculum coverage. If we calculated reliability in a manner that failed to attend to the structure, the phenomena we have counterbalanced against (e.g., a preference for the left or right side in some users) would appear as a form of unreliability.

If we were using a sum or mean score, we could use stratified alpha. However, our scoring algorithm is a form of least squares regression. Previously in the literature, split-half reliability has been used with estimates of the Weber fraction for numerosity (Dietrich et al., 2016). Split-half reliability, with the Spearman-Brown prophecy formula for a full-scale estimate, works for any algorithm that can calculate a score from half the data, and have that score be an unbiased estimator of the full-scale score; specifically, for Spearman-Brown, the full-scale score is the unweighted sum or average of the half-scale scores. The former includes our least squares algorithm, and the latter is a linear approximation.

While stratified alpha is more commonly referenced in our discipline than stratified stratified sampling of items for split-half, very often, the latter does work. We computed split-half reliability for the Weber fractions we obtained from each Magnitudes subtest, based on constrained halves that maintained the counterbalancing rules.

4 Composite Scores: The Story of Balloons

A relatively common challenge in estimating the reliability of scores used in employment testing and organizational development is that many of the scores, as used, are actually linear composites, or more complex functions, of multiple component measures. Composite scoring is done to reflect the task complexity and multifaceted nature of most jobs. For example, an employer may wish to allow candidates or employees to compensate for one area of weakness with another area of strength.

The pymetrics system has multiple layers of composite scoring (especially if taken broadly to include more functional forms than linear combinations). Not only are predictions of job fit made based on machine learning models using dozens of individual measures as inputs, some of those individual measures result from scoring functions that can be themselves interpreted as composites. Two types of these are described here; the first, simpler type includes several of the measures produced by the game Balloons.

Balloons is an implementation of the Balloon Analogue Risk Task (BART) (Fig. 4), which is used to measure risk-taking behavior (Lejuez et al., 2002; Lauriola

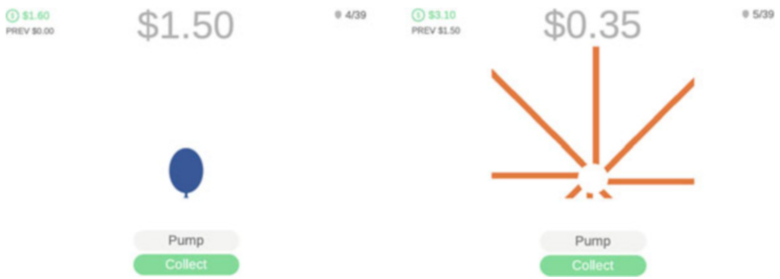


Fig. 4 *Balloons. Left:* Pumping up a blue balloon. This balloon is slowly expanding to fill the white space after thirty pumps, each good for $0.05. *Right:* Oops! This orange balloon popped after seven pumps. The user will *not* receive $0.35

et al., 2014; Charness et al., 2013). It taps into risk propensity and learning under uncertainty.

As a user, you pump up a balloon as much as you can before it pops. If you stop and it hasn't popped, you get paid game money for how much you pumped it up. If you pop the balloon, you get nothing. The popping happens at random within a specified distribution, not at the same number of pumps each time. There are three colors with different pop distributions. Each user is presented thirteen balloons of each color, giving the opportunity to learn over time.

High risk propensity in the BART has been found to be significantly associated with numerous real-world behaviors, including effective workplace maverickism (Gardiner & Jackson, 2012), propensity for gambling (Lejuez et al., 2002, 2003), and novel task learning speed (Humphreys et al., 2013). Low risk propensity, that is, caution, has been associated with more conservative workplace behavior and higher positions of power (Maner et al., 2007). Different job functions and types call for different levels of risk propensity (Nicholson et al., 2005), yet risk propensity remains stable within individuals (Josef et al., 2016). This combination makes it useful to pymetrics for assessing job fit.

The pymetrics implementation of Balloons produces several measures of change between the first and last few trials. Do you become more confident, and increase the number of pumps before you collect your money? Do you figure out where you need to stop to avoid popping so many balloons? Do you get faster or slow down at pumping each balloon? These change scores are computed as differences between ordinary mean scores. This, then, begins our discussion of scores that are derived from multiple simpler scores: composite scores.

Fortunately for the case of Balloons, reliability of difference scores is a solved problem. There is a straightforward formula for the calculation of reliability for difference scores, as given by Feldt and Brennan (1989):

$$\rho_{DD'} = 1 - \frac{\sigma^2_{X_2} * (1 - \rho_{X_2 X_2'}) + \sigma^2_{X_1} * (1 - \rho_{X_1 X_1'})}{\sigma^2_{(X_2 - X_1)}} \tag{3}$$

With this formula, the reliability $\rho_{DD'}$ of the unit-weighted difference score D = X_2 - X_1 can be computed from the reliabilities $\rho_{XX'}$ and variances σ_X^2 of each component measure X, as well as the variance $\sigma^2_{(X_2-X_1)}$ of the difference score itself. The latter works to index how non-redundant the two measures are; taking the difference of two measures that are too closely correlated will result in a low-variance, low-reliability, not-very-useful difference score.

We were able to use this formula for changes in pump speed and pop frequency, and the establishment of a margin before a user expects a balloon to pop, as well as several other difference scores across the pymetrics games. Similar formulas exist, and are presented in the same chapter (Feldt & Brennan, 1989) for other common linear composites such as unweighted or weighted sums of two or more scores.

5 Composite Scores: The Story of Lengths

Not all derived scores, however, are linear composites. Estimates of reliability for derived scores were needed, but in a different way, for the game called Lengths.

Lengths is adapted from the Probabilistic Reward Task (PRT), which was developed to measure reward responsiveness and reinforcement learning (Pizzagalli et al., 2008, 2005). In Lengths, a user is asked to distinguish between two slightly different emoji faces, presented sequentially over seventy trials (Fig. 5). Both faces are variations on the still-face emoji, presented in white lines on a dark blue background for contrast. In the "long mouth" face, the flat-line mouth is a few pixels longer than in the "short mouth" face. The difference between faces is subtle, designed to be a Just Noticeable Difference under reasonable play conditions.

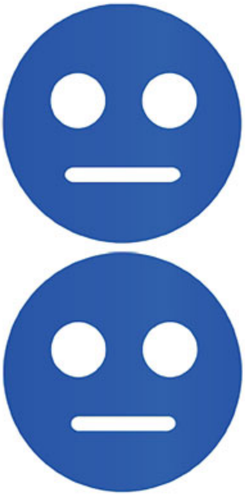


Fig. 5 *Lengths.* Long and short mouth face, enlarged

Users receive intermittent rewards for correct identification, but the reward schemes are different for the two faces, and one has a higher expected value; it is better for earning game money. The higher expected value emoji face is denoted “rich” and the lower expected value condition is denoted “lean.” There are two alternate forms with different reward patterns associated with each face, as well as different sequences of “long mouth” and “short mouth” emoji faces.

The primary measure of Lengths is a measure of reward sensitivity, operationalized as response bias. Do you favor the high-reward face when you’re not sure which one you’re looking at? Do you default to it? And by how much?

The formula for computation of response bias is adapted from earlier literature. Pizzagalli et al. (2008) used the following formula:

$$b = 0.5 * log(\frac{C_{rich} * I_{lean}}{C_{lean} * I_{rich}}) \tag{4}$$

In this formula, the C and I elements are counts of correct and incorrect identifications in each condition. The Pizzagalli formula is equivalent to a difference of log odds ratios:

$$b = 0.5 * (log(\frac{C_{rich}}{I_{rich}}) - log(\frac{C_{lean}}{I_{lean}})) \tag{5}$$

However, in rare cases, this formula can lead to division by zero or taking the logarithm of zero. In order to prevent such cases in our large operational setting, as well as to more conservatively score extreme observed performance (e.g. “lucky streaks” or forms of inattentive responding), each observed response count was increased by one-half response, a simple projection of what might have occurred if the game were extended. (A more sophisticated approach might have been the use of a Bayesian prior; however, the practical effect would be similar.)

$$b = 0.5 * log(\frac{(C_{rich} + 0.5) * (I_{lean} + 0.5)}{(C_{lean} + 0.5) * (I_{rich} + 0.5)}) \tag{6}$$

There remains a simpler formulation as a difference of log odds ratios, albeit with the half-response adjustment included.

$$b = 0.5 * (log(\frac{(C_{rich} + 0.5)}{(I_{rich} + 0.5)}) - log(\frac{(C_{lean} + 0.5)}{(I_{lean} + 0.5)})) \tag{7}$$

It is worth noting that these scoring formulae give relative emphasis to the less-common incorrect identifications. (They are less common in that individuals perform above chance; they are not rare.) Theoretically speaking, errors are more informative than correct responses.

As far as calculating reliability goes, the response bias measure is derived from two simple sum scores and two transformations of the same sum scores; the correct and incorrect counts for each face must add up to the number of times the face is presented. The reliability of each sum score can be estimated simply. However, getting from those component reliabilities to the reliability of b, the response bias measure, is more complex.

Feldt and Brennan (1989) didn't cover products, or quotients, or logarithms of scores! And for good reason. Only linear composites have consistent reliability across users and scores, even if all the components behave according to classical test theory. In a multiplicative composite, the standard error of measurement contribution for each component score depends on the value of the other component score. There are similarly nonlinear behaviors in the standard error of measurement of proportions, exponents and logarithms.

However, as Feldt and Brennan (1989) did, we can apply standard propagation of error rules to obtain each individual's standard error of measurement in a large and representative sample of our user population. That standard error varies for different individuals. Therefore, as in other cases where conditional standard error of measurement is appropriate, we can calculate the marginal reliability across the population.

Error propagation turned out to be simplest for a weighted sum of two log odds components, although there were equally valid possibilities for calculating composite reliability using the original log-of-quotient-of-products version. Because the frequency of correct responses and the frequency of incorrect responses are linearly dependent—they sum to one—each log odds component is an expression of the form described by y in the equation below, where x is proportion incorrect and c is the half-response.

$$y = log(\frac{(1 - x + c)}{(x + c)}) \tag{8}$$

From this, we can derive a function for the standard error of y in terms of the value of x and the standard error of x.

$$s_y = s_x * (\frac{1}{1 - x + c} + \frac{1}{x + c}) \tag{9}$$

Once the standard errors of the log odds components are obtained, they can be used with ordinary linear composite rules, as in the previous section, to produce a reliability estimate across the two mouth lengths.

Using this method, we obtained, as we had hoped, similar marginal reliability estimates for the two alternate forms, supporting their use in parallel.

6 Simulation and Modeling: The Story of Digits

When no formula seems immediately appropriate, we can sometimes gain insight into the reliability of a measure through process modeling. As an example, here is the story of Digits.

Digits is a visual forward digit span task. Several digits are presented on screen one at a time, after which the user is asked to type in those digits, in order. If the user gets it right, the task is repeated with a sequence one digit longer. If the user gets a sequence wrong, the next sequence has one less digit the next time. On the third error, the game terminates, and the primary score is the length of the longest sequence correctly recalled.

A wide variety of jobs call for a good working memory, the ability to hold information briefly in mind without external assistance. Importantly to pymetrics, memory span tasks have been found to show smaller differences between demographic groups than more general measures of cognitive ability (Verive & McDaniel, 1996); it is thus more feasible to create a fair composite score that includes digit span, when it is demonstrated to be job-relevant.

Digit span measures have been around for more than a century, and there are numerous established methods in the literature for computing their reliability, each dependent on the specifics of the task. Non-adapting versions of the digit span task (and other related memory span tasks) are amenable to estimation of split-half reliability or Cronbach's alpha (Waters & Caplan, 2003). Sometimes digit span reliability is obtained just by having some or all users do the task twice, and then correlating the results. However, in an effort to respect our users' time, we use a simple adaptive form and only present the task once.

In order to explore how error of measurement arises and manifests in Digits, we simulated user behavior using a simple item response model that has an error term built in. The model was a two-parameter normal ogive model, with a mean drawn from a distribution that approximates our observed scores, and a constant standard deviation across simulees, spans and trials.

While the standard deviation of the digit span score was observable in our data, the standard deviation parameter of the item characteristic curve (a transformation of item discriminability) was not directly observable. The simulation setup allowed us to test a range of values for the parameter, and map the effects on observable individual and population metrics. One such metric was the difference in length between the highest span with a correct response and the last span tested, which by definition got an incorrect response. The possible values are -1, 0 and 1, except in the rare case where a user enters no correct values, usually due to technical problems.

The objective of the simulation, then, was to relate the standard deviation parameter of the item characteristic curve to the proportions of each of the three possible last span deltas, and from that function, map the actual observed proportions back to a standard deviation parameter, or at least a small range of plausible parameter values. The standard deviation parameter can be translated back into a reliability estimate, using the concept that a reliability $\rho_{XX'}$ is the proportion of score variance σ_X^2 not attributable to error of measurement.

$$\rho_{XX'} = \frac{(\sigma_X^2 - \epsilon^2)}{\sigma_X^2} \tag{10}$$

Because the stopping rule depends on a single (third) error, the reliability of the measure is effectively the reliability of a single item response.

The assumption that the standard deviation parameter (or item discriminability) holds constant across all individuals and sequence lengths is important. There is no reason from the theory of working memory to believe it must be constant; however, it is necessary to obtain a constant estimate of reliability across all users and scores on Digits. Therefore, the simple model was used.

The form of the simulation study allowed for different approaches to estimating the standard deviation parameter. Optimization methods such as maximum likelihood could be used to obtain a single best value. First, however, we chose to plot the simulation results, in order to visually assess the range of plausible values, as well as checking to see that the model was plausible at all given the actual triplet of observed last span delta frequencies.

The simulation results, and the observed user proportions, are plotted in Fig. 6. If the model were perfect, the top, middle and bottom pairs of solid and dashed lines should cross at the same left-right location. They don't. That's a sign of model misfit. In particular, the model doesn't account for the handful of users with technical problems, and also under-predicts a small percentage of users that appear to give up at some point, whether due to frustration or interruptions. These few users have all three sequential errors right at the end, and often the kind of error that suggests not trying (e.g., blank, repetitive or very brief responses).

The misfit isn't severe. There is a plausible range of values defined by the places where the three pairs of lines do cross, and a modest amount of difference in the observed and predicted probabilities through most of that range. That whole

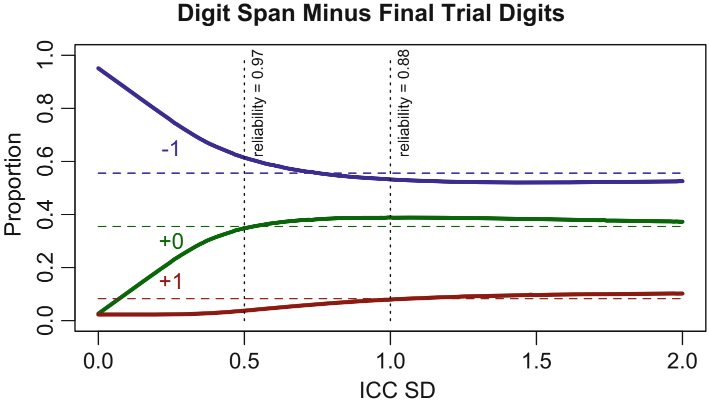


Fig. 6 *Digits:* Difference between two span scores as a function of the standard deviation of the item characteristic curve of recalling an individual sequence of digits. Solid lines represent the simulated frequencies of each last span delta according to the standard deviation parameter of the item characteristic curve for a single trial of recalling a digit sequence. Dashed horizontal lines represent observed frequencies in a population of actual users. Dashed vertical lines show a region of plausible values for the parameter, with the corresponding measure reliability noted

plausible range is consistent with values obtained in prior literature for the reliability of similar digit span tasks (Waters & Caplan, 2003). On the other hand, out toward the right of the plot, all of the curves flatten out, and don't change much. This suggests that neither eyeballing nor optimization is ever going to come to a highly certain and replicable result here. Nor is it likely to be apparent whether conditional standard error of measurement is needed.

For now, we choose to use the most conservative estimate of reliability in the plausible range, with a standard deviation of 1 and a reliability of 0.88.

7 Conclusion

As technology allows measurement to grow more complicated, integrated, and comprehensive, reliability becomes more difficult to estimate, but not less relevant. As with other forms of technical quality assurance, we must innovate to keep up with innovation in scoring, and rely on creativity to keep up with creativity in measurement. I hope these anecdotes provide some inspiration to you, my readers, when your own measures are up to their own shenanigans. Good luck to all of you!

Acknowledgments I would like to thank Frida Polli, founder of pymetrics. I would also like to thank Lewis Baker and Jackson Dolphin for the stories behind the games.

References

Baker, L., & Thissen-Roe, A. (2021). Differences in symbolic and non-symbolic measures of approximate number sense. In M. Wiberg et al. (Eds.), *Quantitative psychology* (pp. x-x). New York: Springer.

Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin, 140*(4), 980–1008.

Charness, G., Gneezy, U., & Imas, A. (2013). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior and Organization, 87*, 43–51.

Dehaene, S. (2007). Symbols and quantities in parietal cortex: Elements of a mathematical theory of number representation and manipulation. In P. Haggard & Y. Rossetti (Eds.), *Attention and Performance XXII Sensorimotor Foundations of Higher Cognition* (pp. 527–574). Cambridge: Harvard University Press.

Dietrich, J. F., Huber, S., Klein, E.,Willmes, K., Pixner, S., & Moeller, K. (2016). A systematic investigation of accuracy and response time based measures used to index ans acuity. *PLoS One*, *11*(9), e0163076.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 105–146). New York: American Council on Education/Macmillan.

Gardiner, E., & Jackson, C. J. (2012). Workplace mavericks: How personality and risk-taking propensity predicts maverickism. *British Journal of Psychology, 103*(4), 497–519.

Humphreys, K. L., Lee, S. S., & Tottenham, N. (2013). Not all risk taking behavior is bad: Associative sensitivity predicts learning during risk taking among high sensation seekers. *Personality and Individual Differences, 54*(6), 709–710.

Hunt, E. (2007). *The mathematics of behavior*. New York, NY: Cambridge University Press.

Josef, A. K., Richter, D., Samanez-Larkin, G. R., Wagner, G. G., Hertwig, R., & Mata, R. (2016). Stability and change in risk-taking propensity across the adult lifespan. *Journal of Personality and Social Psychology, 111*(3), 430–450.

Lauriola, M., Panno, A., Levin, I. P., & Lejuez, C.W. (2014). Individual differences in risky decision making: A meta-analysis of sensation seeking and impulsivity with the balloon analogue risk task. *Journal of Behavioral Decision Making, 27*(1), 20–36.

Lejuez, C.W., Aklin, W. M., Zvolensky, M. J., & Pedulla, C. M. (2003). Evaluation of the balloon analogue risk task (bart) as a predictor of adolescent real-world risk-taking behaviours. *Journal of adolescence*, *26*(4), 475–479.

Lejuez, C.W., Read, J. P., Kahler, C.W., Richards, J. B., Ramsey, S. E., Stuart, G. L.,... Brown, R. A. (2002). Evaluation of a behavioral measure of risk taking: the balloon analogue risk task (bart). *Journal of Experimental Psychology: Applied*, *8*(2), 75–84.

Maner, J. K., Gailliot, M. T., Butz, D. A., & Peruche, B. M. (2007). Power, risk, and the status quo: Does power promote riskier or more conservative decision making? *Personality and Social Psychology Bulletin*, *33*(4), 451–462.

Nicholson, N., Soane, E., Fenton-O'Creevy, M., & Willman, P. (2005). Personality and domain-specific risk taking. *Journal of Risk Research*, *8*(2), 157–176.

Pizzagalli, D. A., Iosifescu, D., Hallett, L. A., Ratner, K. G., & Fava, M. (2008). Reduced hedonic capacity in major depressive disorder: Evidence from a probabilistic reward task. *Journal of Psychiatric Research*, *43*(1), 76–87.

Pizzagalli, D. A., Jahn, A. L., & O'Shea, J. P. (2005). Toward an objective characterization of an anhedonic phenotype: a signal-detection approach. *Biological Psychiatry*, *57*(4), 319–327.

Price, G. R., Palmer, D., Battista, C., & Ansari, D. (2012). Nonsymbolic numerical magnitude comparison: Reliability and validity of different task variants and outcome measures, and their relationship to arithmetic achievement in adults. *Acta Psychologica*, *140*, 50–57.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*, 54–67.

Thissen-Roe, A., & Baker, L. (2021). Estimating approximate number sense (ANS) acuity. In M. Wiberg et al. (Eds.), *Quantitative psychology* (pp. x–x). New York: Springer.

Treadway, M. T., Buckholtz, J. W., Cowan, R. L., Woodward, N. D., Li, R., Ansari, M. S., & Zald, D. H. (2012). Dopaminergic mechanisms of individual differences in human effort-based decision-making. *Journal of Neuroscience*, *32*(18), 6170–6176.

Treadway, M. T., Buckholtz, J. W., Schwartzman, A. N., Lambert, W. E., & Zald, D. H. (2009). Worth the 'EEFRT'? The effort expenditure for rewards task as an objective measure of motivation and anhedonia. *PloS One*, *4*(8), 1–9.

Verive, J. M., & McDaniel, M. A. (1996). Short-term memory tests in personnel selection: Low adverse impact and high validity. *Intelligence*, *23*, 15–32.

Waters, G. S., & Caplan, D. (2003). The reliability and stability of verbal working memory measures. *Behavior Research Methods, Instruments and Computers*, *35*(4), 550–564.

Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J.,... Polli, F. (2021). Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Chicago, IL (pp. 666–677).