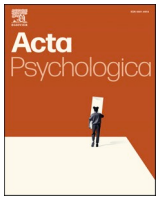
Acta Psychologica 253 (2025) 104703

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy



Risky actions: Why and how to estimate variability in motor performance

John M. Franchak [a,*], Christina M. Hospodar [b], Karen E. Adolph [b]

[a] Department of Psychology, University of California, Riverside, United States of America
[b] Department of Psychology, New York University, United States of America

ARTICLE INFO

Keywords:
Affordances
Motor performance
Risk estimation
Risk perception
Psychophysics
Psychometric

ABSTRACT

We describe the difficulties of measuring variability in performance, a critical but largely ignored problem in studies of risk perception. The problem seems intractable if a large number of successful and unsuccessful trials are infeasible. We offer a solution based on estimates of task-specific variability pooled across the sample. Using a dataset of adult performance in throwing and walking tasks, we show that mischaracterizing the slope leads to unacceptably large errors in estimates of performance levels that undermine analyses of risk perception. We introduce a "pooled-slope" solution that approximates estimates of individual variability in performance and outperforms arbitrary assumptions about performance variability within and across tasks. We discuss the advantages of objectively measuring performance based on the rate of successful attempts—modeled via psychometric functions—for improving comparisons of risk across participants, tasks, and studies.

1. Introduction

Risk perception is critical for action planning across the lifespan—a baby deciding whether to step off the couch, a teen deciding when to cross a busy street, or an elderly person deciding if they can walk over slanting ground. But how do you know if an action is safe or risky, easy or difficult? Actors—like the researchers observing them—estimate risk based on the likelihood of successful performance combined with the penalty for errors (versus rewards for success). This paper focuses on how the likelihood of successful versus failed performance relates to risk: The lower the probability of success, the greater the difficulty and potential risk. Because bodies and skills vary widely, motor performance varies widely—even among people of the same age performing the same task (Franchak, 2019; Ishak et al., 2014; O'Neal et al., 2016; Warren & Whang, 1987; Wilmut & Barnett, 2011). Thus, researchers must measure performance for each participant rather than assume it a priori. That is, to denote people's decisions as "accurate," "cautious," "reckless," and so on, researchers must know the likelihood of success for a particular person attempting a particular action.

1.1. A psychophysical approach to motor performance

Optimally, researchers parametrically vary environmental units to determine the actual success level for each participant in each task (e.g., smaller to larger couch heights for each baby, slower to faster traffic flows for each teen, shallower to steeper slants for each elderly person). A psychophysical approach to modeling motor performance facilitates comparisons across parametric task variations because it reduces performance over repeated trials to two parameters. The *threshold* parameter reflects the environmental unit where the likelihood of success is 50 %. And the *slope* parameter reflects performance variability—that is, how rapidly the likelihood of success decreases from ~100 % (nearly always safe) to ~0 % (nearly always risky) with change in environmental units (Franchak, Adolph, 2014a).

Fig. 1A and B show examples of two infants' attempts to walk over bridges varying in width. Of course, narrower bridges were riskier (babies were more likely to fall) and wider bridges were safer (babies were more likely to succeed). The psychometric functions quantify the likelihood of success across all possible environmental units by specifying the bridge widths where each infant's success rate approached 0 % and 100 %. Thus, the difference in thresholds for baby A (16.8 cm) and baby B (22.0 cm) shows that baby A could walk over narrower bridges than baby B, meaning that a 22 cm bridge was more difficult for baby B than for baby A.

However, motor performance is a continuous measure such that each bridge width (including interpolated widths that were not tested) is associated with its own likelihood of success for each participant. The larger slope (flatter curve) for baby A reflects more gradual change in the likelihood of success across bridge widths, and the smaller slope (steeper curve) for baby B reflects more rapid changes in the likelihood

* Corresponding author.
E-mail address: franchak@ucr.edu (J.M. Franchak).

https://doi.org/10.1016/j.actpsy.2025.104703
Received 19 August 2024; Received in revised form 13 December 2024; Accepted 7 January 2025
Available online 21 January 2025

of success. These differences in performance variability mean that the more variable baby A achieved 75 % success at 2.1 cm larger than their threshold, whereas the more consistent infant B achieved 75 % success at only 0.2 cm larger than their threshold. Thus, accurate estimation of motor performance (and by extension, risk level) across parametric variations in a task requires robust estimates of both the threshold and the slope.

1.2. The importance of accurate estimates of performance variability for comparing risk across tasks, people, and sessions

Researchers—like everyone else—frequently make crude, a priori assumptions about risk. Walking over a wider bridge is safer than walking over a narrow one (Kretch, Adolph, 2013b), descending from a high platform is safer than stepping over a low drop-off (Kretch, Adolph, 2013a), walking through a wide doorway is safer than squeezing through a narrow one (Comalli et al., 2013; Franchak, 2019, 2020; Franchak, Adolph, 2014b; Wilmut & Barnett, 2011), crossing a street is safer when the time gap between cars is longer (O'Neal et al., 2016; Plumert et al., 2007), and so on. Albeit true, such simple a priori assumptions about relative difficulty (narrower bridges and doorways are generally more difficult than wider bridges and doorways, etc.) is not the same thing as quantifying the likelihood of success. The latter requires empirical measurement of motor performance.

To quantify performance, researchers frequently compare people's abilities based on psychophysical estimates of their thresholds. For example, some 14-month-old babies can safely walk over a 12-cm wide bridge, whereas other 14-month-olds display thresholds of 30 cm (Kretch, Adolph, 2013b). Some 18-month-olds can safely walk down 28-cm high drop-offs, whereas other 18-month-olds display thresholds of 2 cm (Karasik et al., 2016; Kretch, Adolph, 2013a). Some 18-month-olds can walk down 40° ramps, whereas other 18-month-olds display thresholds of 12° (Tamis-LeMonda et al., 2008). Individual differences in participants' abilities within a task mean that participants have different likelihoods of success given the same environmental challenge.

However, environmental units like bridge widths, drop-off heights, and ramp degrees vary continuously meaning motor performance also varies continuously. A common strategy for comparing performance at multiple environmental units is to center each participant's data around their threshold and then compare performance at units smaller or larger than the threshold (Adolph, 1995; Franchak & Adolph, 2012; Kretch, Adolph, 2013a, 2013b; Plumert, 1995). However, this strategy lacks precision if individual slope estimates differ, as in infants A and B in Fig. 1. In other words, bridge widths 5 cm narrower than threshold, drop-offs 5 cm higher than threshold, and ramps 5° steeper than threshold entail very different difficulties for participants with flatter or steeper psychometric functions. Thus, to compare performance across units relative to threshold, researchers require accurate slope estimates.

Moreover, when environmental units differ across tasks (as in centimeters of bridge width versus degrees of slant) or hold different functional meanings (as in centimeters of bridge width versus centimeters of drop-off height), cross-task comparisons are so compromised as to be nearly meaningless. That is, there is no reason to assume that walking over a bridge 5 cm narrower than threshold is just as difficult as walking down a ramp 5° steeper than threshold or even over a drop-off 5 cm higher than threshold. Cross-task comparisons would be feasible if performance were equated based on a unit-less metric such as standard deviations from threshold (like a *Z*-score). However, as in within-task comparisons of performance, cross-task comparisons using such a unit-less metric require accurate slope estimates.

1.3. Challenges in estimating performance variability

A well-known maxim in psychophysics is that estimating a slope parameter is more challenging than estimating a threshold parameter (Wichmann, Hill, 2001a, 2001b). The challenge arises for several related reasons. First and foremost, slope estimates require more trials than threshold estimates because the same threshold can result from an infinite number of slopes (imagine a family of psychometric functions centered around a single threshold). And the more variable the performance, the more trials are needed per participant. Because thresholds vary widely among individuals, researchers use adaptive methods, such as staircase procedures (Cornsweet, 1962; Kingdom & Prins, 2010), to minimize the number of trials to estimate threshold and to place trials where they will be maximally informative (Franchak, Adolph, 2014a). The more trials needed to estimate the threshold, the fewer trials are left to fine-tune the slope estimate. Thus, some studies in visual psychophysics collect hundreds or thousands of trials per participant per task to ensure robust slope estimates (e.g., Strother & Kubovy, 2006).

Second, each trial comes at a cost. Immense numbers of trials are feasible in psychophysical studies of adult visual perception, where each trial lasts only a few seconds and a seated observer responds with a mere button press. But not so for many studies of motor action. Walking through doorways (Franchak, 2020; Franchak, Adolph, 2014b; Wagman & Malek, 2007; Yasuda et al., 2014) or stepping over barriers (Comalli et al., 2017; Snapp-Childs & Bingham, 2009) require more effort and

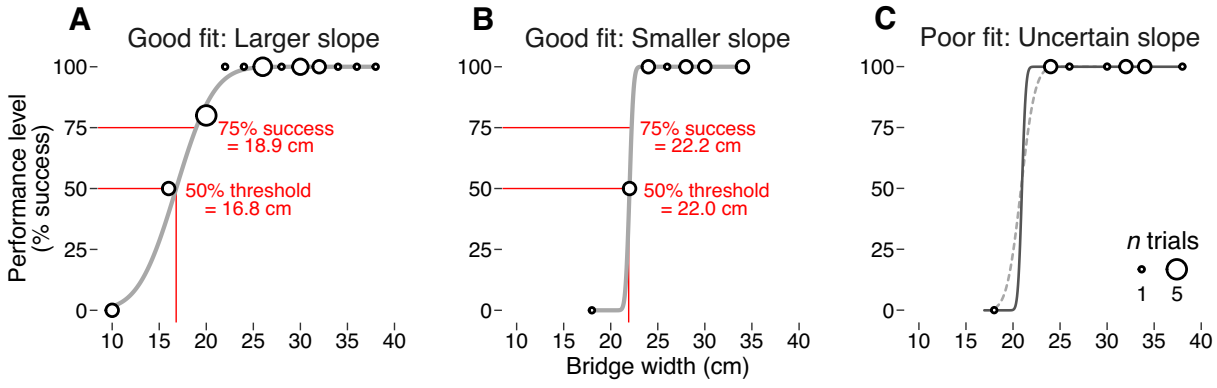


Fig. 1. Examples of psychometric functions for data from three infants walking over bridges in Kretch and Adolph (2013b). Each graph shows the percentage of successful trials (y-axis) at various bridge widths (x-axis). Symbol size is scaled to the number of trials collected at each bridge width. Gray lines are the individual psychometric functions. (A) Good curve fit with relatively large slope (flatter function). (B) Good curve fit with relatively smaller slope (steeper function). Red reference lines show the thresholds (environmental unit with 50 % success rate). Difference between thresholds and 75 % success rate illustrate that performance depends on the size of each infant's slope. When the slope is large as in (A), the difference is 2.1 cm, but when the slope is small as in (B), the difference is only 0.2 cm. (C) Dataset with only a single failure leading to an uncertain slope—both smaller (i.e., black line) and larger (i.e., gray dashed line) slope estimates are possible fits to the data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

time than a button press. And trials are even more costly in tasks like swinging on monkey bars (Cole et al., 2013), leaping over gaps (Day et al., 2015), and climbing up rock walls (Hacques et al., 2021). The cost per trial is exaggerated in infants and children, who become fussy and noncompliant, requiring many easy trials to stay motivated. And the cost is greater for participants who tire quickly, such as infants and elderly people. The higher the cost, the fewer trials can be collected in total.

Third, trials must be informative to fit the psychometric function. That is, each participant must contribute both successful and unsuccessful trials along the inflection of the curve. Trials far on the tails of the curve are less informative because researchers can assume ~100 % success on easy trials such as a 100-cm wide bridge, 0-cm high drop-off, or a 0° slope and they can assume ~0 % success for a 1-cm wide bridge, 200-cm high drop-off, and 70° ramp. Collecting data along the inflection of the curve is more informative, but only if there are enough trials to accurately estimate the performance level. With only 1 trial, the outcome can be either 0 % or 100 %, with 2 trials, the outcome can be 0 %, 50 %, or 100 %, and so on. With many trials at a unit, it is possible to precisely measure how small variations in an environmental unit result in subtle differences in performance. In some cases, the researcher can present multiple trials at a unit and ask participants to attempt each one (Franchak, 2017; Hospodar et al., 2023; Labinger et al., 2018).

However, in many cases, eliciting multiple attempts at a difficult environmental unit is not possible due to participant compliance or safety. For example, elderly adults were reluctant to fall from a narrow ledge, despite the presence of a spotter, so participants refused to attempt to walk over ledges where they thought they would fall—resulting in many successes and few or no failures (Comalli et al., 2013). Babies become frustrated when attempting difficult or impossible increments such as walking over narrow bridges (Fig. 1), down steep ramps, over high drop-offs, or under low barriers, resulting in too few trials at difficult performance levels, and some babies never attempt increments where they will fail (Franchak & Adolph, 2012; Kretch, Adolph, 2013a; Rachwani et al., 2022; Tamis-LeMonda et al., 2008). The consequence of insufficient numbers of unsuccessful trials are poor psychometric fits that mischaracterize the variability. Fig. 1C shows an example where the baby contributed only one unsuccessful trial—leading to an uncertain estimate of the slope that may not characterize the infant's true performance variability. Both the smaller (black line) and larger slope (dashed gray line) are potential fits, but more data would be needed to differentiate the possibilities. As a consequence, lack of trials at difficult risk levels makes it challenging to estimate the slope.

1.4. Current study

We present a new psychophysical method for estimating performance variability that meets the challenges inherent in comparing physical risk across tasks with different units or functional meanings, across participants with different performance variability due to skill differences, and across sessions with different performance variability due to learning or development. We tested the effectiveness of our method for dealing with the additional problems introduced by small numbers of imbalanced trials as is typical in studies with infants, children, and elderly adults or in tasks where motor performance is especially costly.

We used an existing dataset of adults throwing a beanbag through doorways and walking sideways through doorways of varying widths (Hospodar et al., 2023). Although doorway width was measured in the same units (cm) in both tasks, throwing a beanbag and walking through doorways have different functional meanings in terms of difficulty and risk. Thus, comparisons across tasks should be based on performance level in % success rather than absolute environmental units, but this would require accurate estimates of the slope. Critically, each participant contributed a relatively large number of trials (75 per task) with multiple trials at easy and difficult environmental units. Thus, we treated the individual fits to each participant's 75 trials of data as the "ground truth" estimates for performance variability in each task. We then tested how alternative estimates of performance variability compare to the ground truth estimates. Based on individual slope estimates from psychometric functions fit through each participant's data in each task, the prior work revealed differences in performance variability across participants within each task, and showed that performance variability across participants was greater for throwing compared with walking.

The throwing/walking dataset was ideally suited to address four related aims. First, we tested whether differences in performance variability lead to differences in the robustness of slope estimates. More trials are needed to estimate the slope of a psychometric function when variability is larger (curve is flatter) than when it is smaller (curve is steeper). Thus, we compared the size of the confidence intervals around slope estimates in the throwing task to the confidence intervals in the walking task while keeping the number of trials constant across tasks. We predicted larger confidence intervals (indicating inconsistent slope estimates) for throwing than for walking.

Second, we tested the consequences of misrepresenting performance variability (slope estimates), an issue that arises when researchers assume risk level a priori rather than measure it empirically. Thus, we calculated a new metric, a "performance estimation error score," that reveals errors in performance estimation (estimated % success) at every point along the psychometric function. Specifically, we compared the ground truth slope estimates (based on individual participant fits using all 75 trials) in the throwing and walking tasks with the misrepresented slope estimates researchers would obtain if they assumed identical performance variability across the two tasks—as is typical in risk-perception research (e.g., Franchak, 2020; Franchak et al., 2012; Plumert, 1995; Yasuda et al., 2014). We predicted that misrepresented slopes would lead to unacceptably large performance estimation error scores.

Third, we tested whether data pooled across participants in a task can replace individual estimates of performance variability in studies where participants cannot contribute a sufficient number of trials to estimate individual slopes. Thus, we calculated a "pooled slope" estimate for each task by normalizing each participant's data to their threshold and then fitting a psychometric function to the pooled data. We compared the pooled slope to the individual ground truth slope estimates and to the misrepresented slope estimates. We predicted that the pooled slope provides acceptable estimates of performance for most individual participants.

Finally, we tested whether the pooled slope is robust for use with datasets where participants contribute only a small number of imbalanced trials, as is common in studies with participants who are loath to fail (e.g., infants, children, and elderly adults). Thus, we simulated such cases by "degrading" the original data to varying degrees to approximate small, imbalanced datasets that lack data at difficult environmental units. Then we compared the performance estimation error (relative to ground-truth risk estimates) in individual fits versus the pooled slope for degraded datasets. We predicted that the pooled slope would provide performance estimation errors less than or equal to errors from individual slope estimates when fitting psychometric functions under the degraded conditions.

2. Method

All analyses in this paper were based on a dataset openly shared on Databrary (databrary.org/volume/1448) and described in Hospodar et al. (2023). The processed data and code to reproduce our analyses are available on OSF (DOI: 10.17605/OSF.IO/WNCBK).

2.1. Procedure for throwing/walking dataset

Thirty participants (18 women, 12 men) aged M = 25.7 years participated in the throwing and walking tasks. Each participant

contributed 75 trials to each task, split approximately evenly between successful and unsuccessful trials, with trials blocked between the two tasks and task order counterbalanced. Participants stood 1.5 m from a doorway that could be adjusted in width from 0 to 74 cm. On each trial, doorway width was determined by an adaptive psychophysical procedure, where width was increased or decreased depending on the participant's success on the previous trial. The procedure was tailored to individual participant's responses so as to minimize the total number of trials while placing trials along the inflection of the psychometric function to be maximally informative.

The edges of the doorway were lined with small bells that jingled if touched. In the throwing task, participants attempted to throw a beanbag through the doorway without the beanbag touching the edges of the doorway. On successful throws, the beanbag passed through the doorway without ringing the bells. In the walking task, participants attempted to walk sideways through the doorway without touching the sides of the doorway. On successful walks, the participant passed through the doorway without ringing the bells.

2.2. Individual threshold and slope estimates in the throwing/walking dataset

Using the *quickpsy* package in R (Linares & López-Moliner, 2016), Hospodar et al. (2023) fit cumulative normal psychometric functions to the success rates for each participant in each task by estimating parameters for the threshold and the slope. Here, we treated the individual psychometric functions as the "ground truth" dataset because they contained a sufficient number of trials (75) to robustly estimate the threshold and slope for each participant. For each participant and task, Hospodar et al. (2023) calculated the success rate as the percent of successful attempts at each doorway width. The success rate estimates the performance level for that participant for a particular doorway width—a success rate of 0 % indicates the action is impossibly difficult and thus riskier, whereas a success rate of 100 % indicates the action is a "sure thing" and thus safer. Each symbol in Fig. 2A shows the success rate for walking (in blue) and throwing (in orange) for each participant. The solid lines in Fig. 2A show the psychometric functions, the white squares show the threshold estimates (the doorway width where likelihood of success was 50 %), and the flatness or steepness of the curves show the slope estimates (the change in likelihood of success with each change in doorway width relative to threshold). Thus, thresholds denote the doorway width where the likelihood of success is 50 %, whereas slopes denote how rapidly performance changes for doorways smaller or larger than threshold.

Overall, thresholds for throwing ranged from 14.2 cm to 27.1 cm (M = 19.2), and thresholds for walking ranged from 29.4 cm to 39.5 cm (M = 34.3). That is, participants could throw beanbags through much narrower doorways than they could walk through, but thresholds varied widely within tasks based on participants' body size and throwing ability. Overall, slopes for throwing ranged from 4.0 cm to 20.9 cm (M = 7.64 cm), and slopes for walking ranged from 0.27 cm to 7.56 cm (M = 2.08 cm). That is, performance changed rapidly around the threshold for the walking task, but performance varied more gradually around the threshold for the throwing task because throwing was less consistently successful. Note that slope estimates of a cumulative Gaussian function are in the same units as the threshold (in this case, centimeters).

3. Results

We report four related sets of results. First, we show that task-specific differences in performance variability lead to differences in the robustness of slope estimates. Second, we show that misrepresentation of performance variability leads to large errors in performance estimation within tasks. Third, we show that pooled slopes provide an acceptable estimate of performance level when used in place of individual psychometric functions. Finally, we show that pooled slopes more accurately estimate performance compared with individual slope estimates when the total number of trials or number of unsuccessful trials is small.

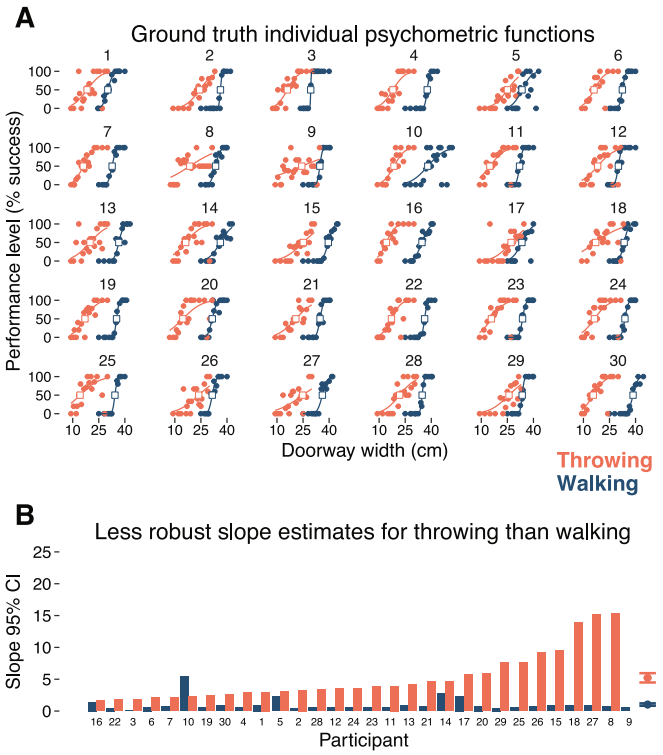


Fig. 2. Data from the throwing/walking task in Hospodar et al. (2023). (A) Individual psychometric functions fit to each participant's data in the throwing (orange) and walking (blue) tasks. Symbols show the percent of successful trials (y-axis) at each doorway width (x-axis). White squares indicate the threshold estimates. (B) Relation between performance variability and robustness of slope estimates. Each bar indicates the size of the 95 % confidence interval for the slope parameter for each participant in each task calculated from bootstrap resampling. Confidence intervals were larger (worse) for throwing compared to walking. Points to the right of the bar graphs indicate group-level means with ±1 SE error bars. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.1. Task-specific differences in performance variability lead to differences in accuracy of slope estimates

With the throwing/walking dataset, we empirically demonstrated a well-known problem in psychophysics—that more trials are needed to estimate the slope when variability is larger (as in throwing) compared to when it is smaller (as in walking). We conducted 1000 parametric bootstraps for each participant's data in each task and then refit the psychometric functions to derive 95 % confidence intervals for the slope estimates. Small confidence intervals denote robust slope estimates across parametric variations in the data whereas large confidence intervals indicate inconsistent slope estimates.

Fig. 2B shows the size of the 95 % confidence intervals for the slope estimate for each participant, ordered by the size of the confidence intervals for the throwing task. One participant's (#9) slope in the throwing task could not be reliably estimated in the majority of simulations and so was omitted. Greater variability in the throwing task resulted in larger confidence intervals for throwing (M = 5.22 cm) compared with walking (M = 1.03 cm) even though each task had the same number of trials. The problem of large confidence intervals (hence, poor slope estimates) is compounded in cases where fewer trials are available, as we describe in the final section of the results.

3.2. Misrepresented performance variability leads to large errors in estimates of performance level

Accurate estimates of performance variability—slope parameters—are critical to characterize the likelihood of success. So we tested the consequences of misrepresenting performance variability as occurs when researchers assume equal variability across tasks. (Note, the same consequences arise if researchers assume that variability is constant across people or across sessions, or if researchers do not measure variability at all.) Thus, to test the accuracy of performance estimates when assuming that the variability of one task is equal to that of another, we swapped slope estimates for throwing and walking, where differences in performance variability were large.

Fig. 3A-B illustrates the consequences of misrepresenting performance variability for one participant (#4 in Fig. 2A). The participant's ground-truth slopes (in gold) show the success rate for doorways larger and smaller than the participant's throwing and walking thresholds. The "swapped slope" estimate (in purple) for the throwing task (Fig. 3A) was the ground-truth slope for the walking task, and vice versa for the walking task. We calculated a "performance estimation error" to quantify the size of the deviations (in performance-level units of % success) between the true success rate in the ground-truth slope model and the success rate in the swapped-slope model. The deviations from the swapped-slope model (vertical purple lines) to the ground-truth slope model show the size of the errors in estimating performance at a few exemplar units (1, 3, and 5 cm larger than threshold). The longer the line, the larger the error.

Fig. 3C shows the range in performance estimation errors for each participant (thin purple lines) across doorway widths. Because the cumulative normal psychometric function is symmetrical around the threshold, the figure shows only positive values up to 5 cm larger than threshold. Exemplar participant #4 is shown in Fig. 3C with a thick purple line; circles indicate the example points at 1, 3, and 5 cm from Fig. 3A-B. Although a few participants had similar slopes for the two tasks (and thus small performance estimation errors when the slopes were swapped), most did not. The average performance estimation errors (thick black lines) are identical for throwing and walking because the slopes were swapped between tasks. The average performance

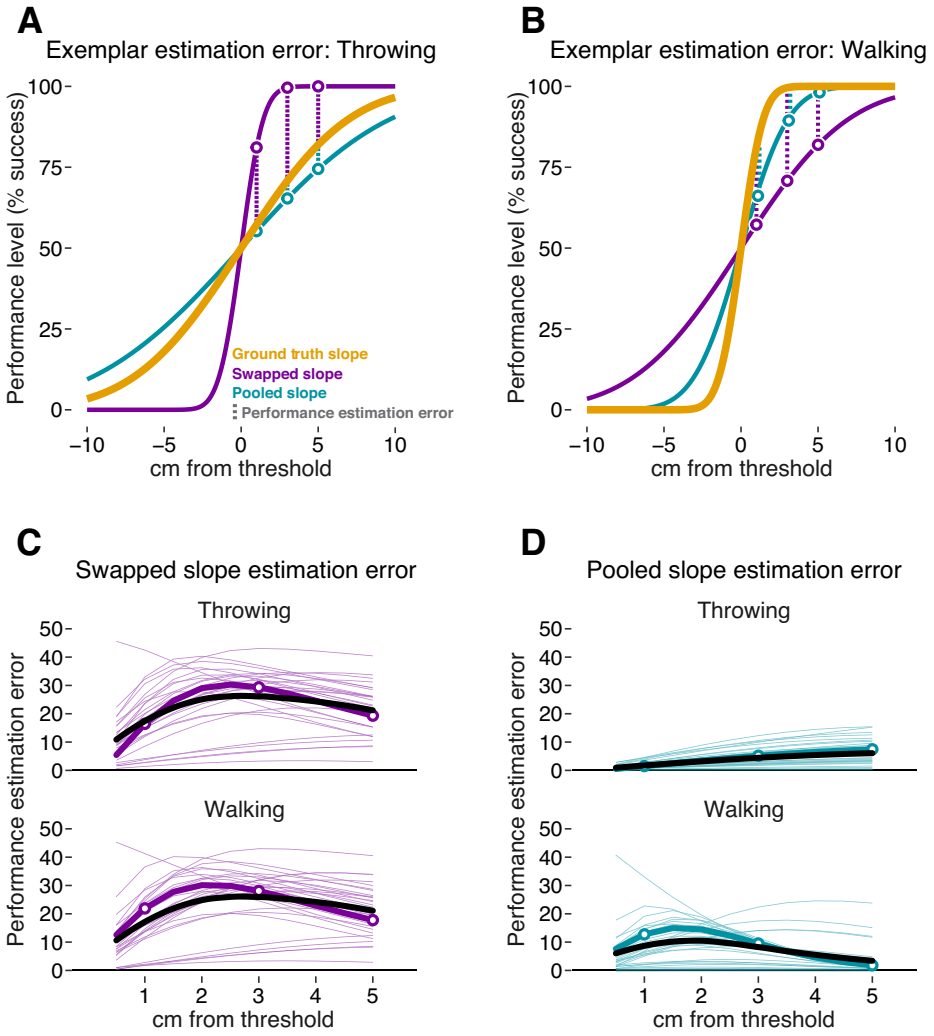


Fig. 3. Performance estimation error (calculated as the absolute difference in percent success compared to ground truth estimates). (A) Throwing and (B) walking. One participant's (#4) percent success (y-axis) at different cm from threshold (x-axis) for the ground truth psychometric function (gold line), the swapped slope psychometric function (purple line), and the pooled slope psychometric function (teal line). The length of the dashed lines in (A) and (B) at 1, 3, and 5 cm show examples of performance estimation error for swapped slopes and pooled slopes. In each case, performance estimation error was larger for swapped slopes than for pooled slopes. (C) Each participant's performance estimation error for throwing and walking for the swapped slope estimates (thin purple lines). (D) Each participant's performance estimation error for throwing and walking for the pooled slope estimates (thin teal lines). White circles and thick purple or teal lines denote the performance estimation error from exemplar participant #4 in (A) and (B). Black lines indicate the mean performance estimation error across participants. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

estimation error was smallest at 0.5 cm from threshold (10.5 % error) and largest at 2.5 cm from threshold (26.1 %). Thus, a large error in misrepresenting the slope parameter results in large errors in estimations of participants' performance levels at units smaller and larger than threshold.

3.3. Pooled slopes provide acceptable estimates of performance level

Robust estimates of the slope are precluded when performance variability is large and/or the total number of trials or number of unsuccessful trials are small. Thus, we present a new procedure to estimate a "pooled slope" for each task, and we tested whether the pooled slope estimate can be used in lieu of individual slope estimates.

The pooled slope combines data across participants in a task. However, because participants had a wide range of thresholds for walking and throwing (Fig. 2A), as is typical in many motor tasks, we first centered each participant's environmental units (doorway width) to their thresholds. Note, threshold estimates based on a smaller number of trials are often robust even when slope estimates are not (Wichmann, Hill, 2001a). Thus, we centered each participant's data by subtracting their threshold from each doorway width such that 0 on the x-axis of Fig. 4A denotes success rate at threshold (50 %). Values smaller than 0 denote increasingly risky doorway widths with a lower likelihood of success; values larger than 0 indicate increasingly safe doorway widths with a higher likelihood of success. With each participant's data centered to a common performance scale, we pooled data across participants in each task. The symbol size in Fig. 4A denotes the number of trials at each unit pooled across participants. The adaptive procedure used to choose doorway widths resulted in more frequent trials near threshold compared with trials at units more distant from threshold.

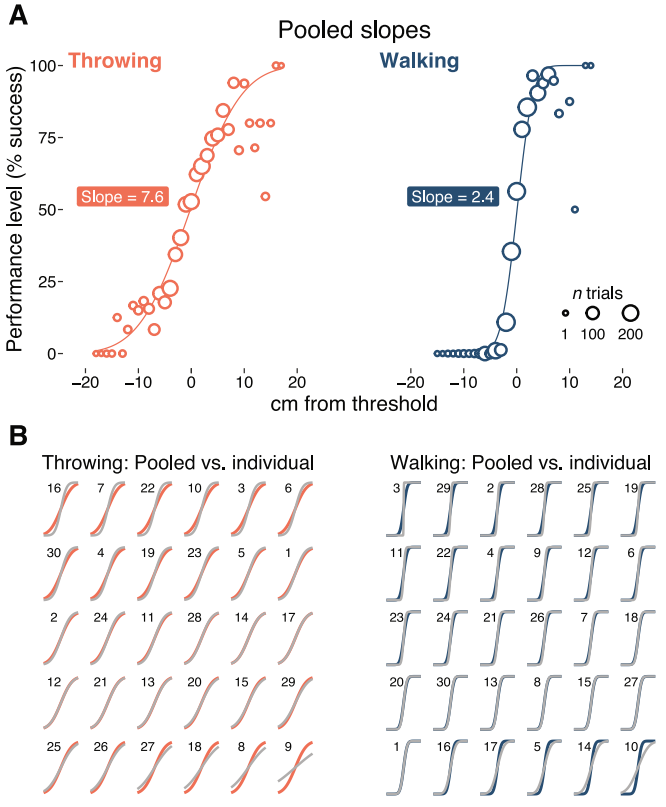


Fig. 4. Pooled slopes for each task. (A) Each participant's data were centered based on their individual thresholds for each task and then rounded to integer units. Each symbol reflects the percent of successful trials in the pooled data after pooling across participants; symbol size was scaled to the number of trials. A single psychometric function in each condition yielded a pooled slope used in subsequent analyses. (B) Pooled slopes (orange and blue curves) relative to each participant's psychometric function for throwing and walking (gray curves). Participants' data were ordered from smallest to largest ground truth slope parameter (i.e., gray curves). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Next, we fit a psychometric function to the pooled data for each task. As with individual psychometric functions, we fit a cumulative normal psychometric function by estimating the threshold and slope using the *quickpsy* package (Linares & López-Moliner, 2016). The solid lines in Fig. 4A show the pooled psychometric functions for each task. The resulting pooled slopes (7.6 cm for the throwing task and 2.4 cm for the walking task) can now be considered as a replacement for individual slope estimates. Fig. 4B-C visualizes the goodness of fit by plotting ground-truth individual slope psychometric functions against the pooled slope psychometric functions.

As with the swapped-slope estimates, we calculated performance estimation errors for the pooled slope estimates. Fig. 3A-B shows the performance estimation error for throwing and walking for the pooled slope (in teal) relative to exemplar participant #4's individual ground-truth slope estimates (in gold). The vertical teal lines show the size of the performance estimation error at 1, 3, and 5 cm. Each line is shorter for the pooled slope compared with the swapped-slope estimate, indicating that the pooled slope is a better estimate. That is, pooled-slope estimates of performance were closer to ground truth compared with swapped slope estimates.

Fig. 3D shows the performance estimation error from pooled slope estimates for each participant in each task. Compared to the swapped slopes in Fig. 3C, the average performance estimation errors for pooled slope estimates were less severe (thick black lines). For throwing, the mean pooled slope errors were smallest at 0.5 cm from threshold (0.9 %) and largest at 5 cm from threshold (6.1 %). For walking, the mean pooled slope errors were smallest at 5 cm from threshold (3.4 %) and largest at 2 cm from threshold (10.5 %). Put differently, the largest averaged error from pooled slope estimates was equal to the smallest averaged error from misrepresenting the slope (10.5 %), showing that pooled slope estimates are preferable to arbitrary slope estimates.

The pooled slope procedure provides a good estimate because each participant's data is centered to a common performance scale (cm relative to threshold) before data are pooled. To illustrate the advantage of centering the data, we created an "uncentered aggregate" for each task by simply calculating the percent of successful trials at each environmental unit across participants. Supplemental Fig. 1A shows the resulting fits when data are aggregated without centering data to a common performance scale. Because participants have varying thresholds in each task, the simple aggregate fits in Supplemental Fig. 1A result in much larger estimates of performance variability—a slope of 12.2 cm for the throwing task (compared with a pooled slope of 7.6 cm) and a slope of 4.5 cm for the walking task (compared with a pooled slope of 2.4 cm). This overestimation happens because between-participant variability in thresholds is (erroneously) being counted in the estimate of within-participant performance variability. As expected, performance estimation errors for the uncentered aggregate fits were larger compared to those for the pooled slope procedure (Supplemental Fig. 1B). Across performance levels ±5 cm from threshold, performance estimation errors for the uncentered aggregate fits averaged $M = 20.6$ % for the throwing task and $M = 12.5$ % for the walking task. In contrast, pooled slopes performance estimation errors averaged only $M = 7.4$ % for the throwing task and $M = 3.6$ % for the walking task across the same performance levels.

3.4. Pooled slopes provide better estimates of performance than individual slope estimates for small numbers of imbalanced trials

Although individual slope estimates are preferable given sufficient data, many test situations allow only a small number of trials and/or a small number of unsuccessful trials. In particular, infants, children, and elderly people can tolerate only a small number of trials and they

sometimes refuse to attempt actions where they will be unsuccessful such as walking over impossibly narrow bridges (Kretch, Adolph, 2013b) or over impossibly narrow ledges (Comalli et al., 2013). Even with healthy, young, compliant adults, some tasks (e.g., arm-leaping on monkey bars), are so arduous or costly that only a small number of trials with a small number of unsuccessful trials are possible (Cole et al., 2013). Thus, we tested whether the pooled slope can provide more accurate estimates than individual slope estimates.

We created infant-inspired datasets by degrading the throwing/walking data in a series of simulations. For each participant in each task, we randomly selected 20 of their 75 trials under three conditions of imbalance: 15 successful and 5 unsuccessful trials, 17 successful and 3 unsuccessful trials, and 19 successful and 1 unsuccessful trial. Hence, we kept the total number of trials constant (20) but varied the number of unsuccessful trials. For each participant, task, and imbalanced-trial condition, we simulated 1000 datasets.

For each simulated dataset, we compared the performance estimation errors of individual slope estimates compared with curves that used pooled slope estimates with individual threshold fits. Fig. 5 shows the average performance estimation error across simulations and participants for each method (individual slopes in brown, pooled slopes in teal). We used 95 % confidence intervals to compare which estimate was most accurate. For example, at 0.5 cm from threshold in the walking task in Fig. 5 (leftmost highlighted points), the performance estimation error was 24.5 % for individual slopes compared with 16.0 % for pooled slope estimates and the 95 % confidence intervals did not overlap. The highlighting shows cases where the pooled slope estimate outperformed the individual slope estimate (non-overlapping confidence intervals where the pooled slope model had smaller errors than the individual slope model) and unshaded areas show where the confidence intervals overlapped.

Individual slope estimates never outperformed pooled slope estimates (confidence intervals always overlapped when performance estimation error was lower for individual slopes compared to pooled slope estimates). The pooled slope estimates were most effective in reducing performance estimation errors in the more variable throwing task for performance levels close to threshold when there were only 3 or 1 unsuccessful trials (Fig. 5). The pooled slope estimates were comparable to individual slope estimates for most performance levels in the walking task regardless of the number of unsuccessful trials, and for the simulations with 5 unsuccessful trials in the throwing task (Fig. 5).

4. Discussion

Using an existing dataset from Hospodar et al. (2023), we showed that performance variability critically influences physical risk because of its influence on the likelihood of successful action. We described a psychophysical procedure to accurately measure performance variability. Hospodar et al. (2023) found that variability of throwing beanbags through doorways exceeded variability for walking through doorways. Consequently, we found here that slope parameter estimates for the more variable throwing task were less robust given the same, relatively large number of trials (75) as the less variable walking task.

Most importantly, we presented a new “pooled slope” procedure that aggregates data across participants in a task. The pooled slope provides acceptable estimates of performance variability (relative to ground truth), whereas a priori assumptions about performance variability (e.g., that two tasks are equally variable) can lead to gross mischaracterization of risk levels. Pooled slopes better approximate performance variability compared to using uncentered, aggregate data (i.e., success rates at each absolute environmental unit calculated across participants). Moreover, the pooled slope procedure mitigates the pitfalls of estimating slope parameters in datasets with too few trials. Pooled slope estimates also outperform individual slope estimates when analyzing datasets that contain only a few unsuccessful trials, such as when studying performance in infants and elderly people (e.g., Comalli et al., 2013; Kretch, Adolph, 2013a).

4.1. From motor performance to risk perception

Of note, the current results focused exclusively on motor performance—whether participants successfully or unsuccessfully performed the target action—rather than on perceptual judgments. However, we argue that researchers must objectively measure performance to study risk perception—that is, whether participants accurately perceive actions as safe or risky where the likelihood of success is known. Imagine the difficulty of studying luminance perception without being able to manipulate or measure the actual luminance of different stimuli! In such a case, results could not be compared across tasks, participants, or sessions.

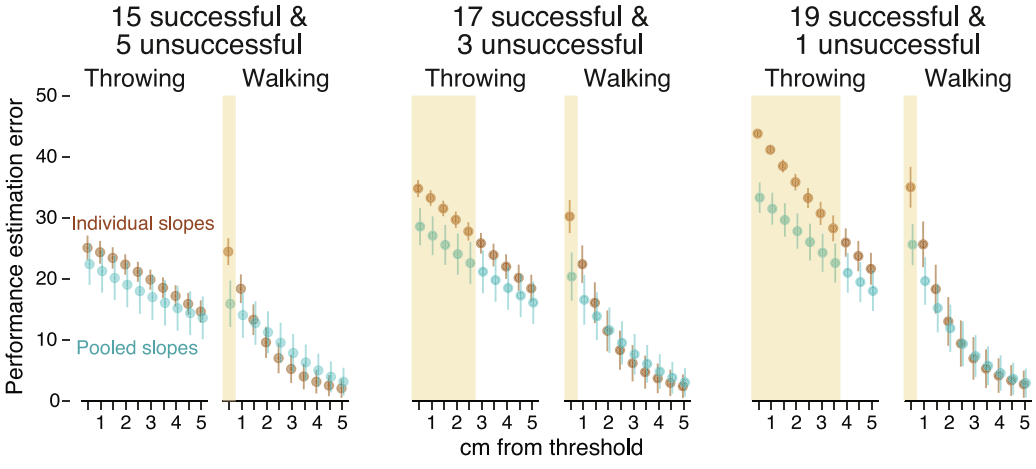


Fig. 5. Performance estimation errors for pooled slopes and ground-truth estimates in infant-inspired datasets with a small number of trials and unsuccessful trials. Average performance estimation error of individual fits to degraded data (brown points) compared with pooled-slope estimations (teal points) with 95 % confidence intervals. Each panel shows a different degraded dataset pulling $n = 20$ random trials from each participant's data in each task: 15 successful and 5 unsuccessful trials (left), 17 successful and 3 unsuccessful trials (center), and 19 successful and 1 unsuccessful trial (right). Yellow shading denotes cases where the pooled slope outperformed individual psychometric functions (non-overlapping confidence intervals). In every non-shaded case, the pooled slope was comparable to individual psychometric functions (overlapping confidence intervals). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Studies of risk perception typically manipulate a single environmental dimension that presumably increases or decreases risk: Narrower doorways increase the risk of entrapment, steeper ramps increase the risk of falling, faster traffic flows increase the risk of collision, and so on (Adolph, 1995; Franchak, 2019; Franchak & Adolph, 2012; O'Neal et al., 2016; Plumert et al., 2007; Tamis-LeMonda et al., 2008). However, possibilities for action depend on the fit between actor and environment (Gibson, 1979; Warren, 1984; Warren & Whang, 1987), and the affordance fit cannot be reduced to only a single parameter. For example, Hospodar et al. (2023) showed that success at throwing beanbags and walking through doorways depends on both threshold and slope parameters; comparisons based only on doorway width are necessarily imprecise. By analyzing performance level rather than doorway width, we can better equate performance (and thus risk) across participants and tasks, and better yet, the focus is on performance rather than one of its many constituent dimensions, such as doorway width. That is, tests of risk perception should be based on performance level in combination with the reward/penalty that results from success/failure, not on doorway width or any other environmental dimension.

4.2. Performance variability and cross-task comparisons

The importance of performance variability in determining risk level means that comparisons among tasks and studies is more fraught than previously acknowledged. For example, Yasuda et al. (2014) and Franchak and Somoano (2018) separately tested whether practice walking through doorways varying in width improves subsequent risk perception. Practice trials in each study were based on metric doorway size relative to threshold without accounting for performance variability. Thus, the actual performance level of those practice trials (which depends on each participant's performance variability) was unknown and might not have been equivalent. Consequently, the accuracy of participants' perceptual judgments may have differed due to unequal practice, but readers would have no way of knowing that.

We suggest that studying risk perception in the language of performance level (% success) is the way forward to allow researchers to compare participants, tasks, sessions, and experiments. The advantage of performance level is that it can describe a wide range of motor tasks using a single metric that abstracts across the multiple actor-environment dimensions that influence performance. Not all metrics have these properties. For example, “pi numbers” relate an environmental dimension (e.g., doorway width) to a body dimension (e.g., shoulder width) as a ratio (Warren & Whang, 1987). However, pi numbers can only account for two dimensions, one about the actor and one about the environment, when other factors might matter (e.g., walking speed, lateral sway of the shoulders). Moreover, pi numbers from one task are not comparable to pi numbers from another task. Whereas the pi number for walking through doorways is 1.3 when relating shoulder width to doorway width, the pi number for walking under barriers is 1.00 to 1.04 when relating participant height to barrier height (Franchak et al., 2012; Stefanucci & Geuss, 2010; van der Meer, 1997). Pi numbers do not equate success rates across tasks like doorway passage and overhead clearance, but performance level in % success allows for meaningful comparisons. Moreover, pi numbers may not be comparable across age groups (older and younger adults) or skill levels (gymnasts and regular folk), even within the same task (Konczak et al., 1992). But if performance were measured in % success units, performance levels could be equated.

However, these advantages can only be realized if researchers can accurately model performance. Prior work showed that a psychophysical curve fitting approach with threshold and slope parameters provides advantages beyond approximating performance into a single critical point, such as the smallest possible doorway (Franchak, Adolph, 2014a). Here, we extend that work to show that researchers have different options for estimating thresholds and slopes depending on the amount of data. In situations where many successful and unsuccessful trials can be collected around the threshold, researchers can fit individual psychometric functions to each participant's data (as in Fig. 2A). However, in situations where such trials are too costly or few, researchers are better off fitting individual thresholds and pooling data across participants to estimate a pooled slope (as in Fig. 4A). Bootstrapped refitting procedures to estimate confidence intervals (as in Fig. 2B) for each parameter can help researchers to understand the precision of their estimates to guide their decisions about the quality of individual versus pooled slopes (Wichmann, Hill, 2001b). User-friendly packages for psychophysical curve fitting, such as *quickpsy* in R (Linares & López-Moliner, 2016), make these approaches tractable for researchers who are unfamiliar with psychophysical methods. Our shared analysis scripts (DOI: 10.17605/OSF.IO/WNCBK) describe each procedure reported in our results.

5. Conclusions

Risk perception is a central phenomenon in studies of motor development (Adolph, 2019), injury prevention (Plumert & Kearney, 2014), and decision making (Dekker & Nardini, 2016). In the laboratory, risk levels in computerized tasks are dictated by the researcher who programmed the experiment (Levy et al., 2010). But physical risk levels in most real-world tasks are beyond the researcher's control because motor performance depends on body-environment relations (Gibson, 1979). Researchers cannot yet model those factors a priori—even in tasks as simple as throwing a bean bag or walking through a doorway. So, we must measure the motor performance of a particular person in a particular task. Improvements in estimating motor performance, such as the pooled slope estimate presented here, allow for better comparisons of risk across people, tasks, sessions, and studies. Such comparisons allow researchers to better understand how motor performance interacts with penalties and rewards as people weigh their motor decisions, and to test safety interventions based on risk perception and physical risk.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.actpsy.2025.104703.

Permission to reproduce materials from other sources

No materials are reproduced from other sources.

CRediT authorship contribution statement

John M. Franchak: Writing – review & editing, Writing – original draft, Visualization, Software, Formal analysis, Conceptualization. **Christina M. Hospodar:** Writing – review & editing, Writing – original draft, Funding acquisition, Data curation, Conceptualization. **Karen E. Adolph:** Writing – review & editing, Writing – original draft, Funding acquisition, Conceptualization.

Ethics approval statement

All study procedures were approved by New York University's IRB (NYU IRB FY2019-3295). All participants gave their informed consent prior to inclusion in the study.

Funding

This research was supported by National Institute of Child Health and Human Development grants to Karen Adolph (R01-HD033486) and Christina Hospodar (F31-HD107999).

Declaration of competing interest

John M. Franchak – I have nothing to declare.
Christina M. Hospodar – I have nothing to declare.
Karen E. Adolph – I have nothing to declare.

Data availability

All analyses in this paper were based on a dataset openly shared on Databrary (databrary.org/volume/1448) and described in Hospodar et al. (2023). The processed data and code to reproduce our analyses are available on OSF (DOI: 10.17605/OSF.IO/WNCBK).

References

Adolph, K. E. (1995). Psychophysical assessment of toddlers' ability to cope with slopes. *Journal of Experimental Psychology: Human Perception and Performance, 21*, 734–750.

Adolph, K. E. (2019). An ecological approach to learning in (not and) development. *Human Development, 63*, 180–201.

Cole, W. G., Chan, G. L. Y., Vereijken, B., & Adolph, K. E. (2013). Perceiving affordances for different motor skills. *Experimental Brain Research, 225*, 309–319.

Comalli, D. M., Franchak, J. M., Char, A., & Adolph, K. E. (2013). Ledge and wedge: Younger and older adults' perception of action possibilities. *Experimental Brain Research, 228*, 183–192. https://doi.org/10.1007/s00221-013-3550-0

Comalli, D. M., Persand, D., & Adolph, K. E. (2017). Motor decisions are not black and white: Selecting actions in the 'gray zone'. *Experimental Brain Research, 235*, 1793–1807.

Cornsweet, T. N. (1962). The staircase-method in psychophysics. *American Journal of Psychology, 75*, 485–491.

Day, B. M., Wagman, J. B., & Smith, P. J. K. (2015). Perception of maximum stepping and leaping distance: Stepping affordances as a special case of leaping affordances. *Acta Psychologica, 158*, 26–35.

Dekker, T. M., & Nardini, M. (2016). Risky visuomotor choices during rapid reaching in childhood. *Developmental Science, 19*(3), 427–439.

Franchak, J. M. (2017). Exploratory behaviors and recalibration: What processes are shared between functionally similar affordances? *Attention, Perception, & Psychophysics, 79*, 1816–1829.

Franchak, J. M. (2019). Development of affordance perception and recalibration in children and adults. *Journal of Experimental Child Psychology, 183*, 100–114.

Franchak, J. M. (2020). Calibration of perception fails to transfer between functionally similar affordances. *Quarterly Journal of Experimental Psychology, 73*, 1311–1325.

Franchak, J. M., & Adolph, K. E. (2012). What infants know and what they do: Perceiving possibilities for walking through openings. *Developmental Psychology, 48*, 1254–1261.

Franchak, J. M., & Adolph, K. E. (2014a). Affordances as probabilistic functions: Implications for development, perception, and decisions for action. *Ecological Psychology, 26*, 109–124.

Franchak, J. M., & Adolph, K. E. (2014b). Gut estimates: Pregnant women adapt to changing possibilities for squeezing through doorways. *Attention, Perception, and Psychophysics, 76*, 460–472.

Franchak, J. M., Celano, E. C., & Adolph, K. E. (2012). Perception of passage through openings cannot be explained geometric body dimensions alone. *Experimental Brain Research, 223*, 301–310.

Franchak, J. M., & Somoano, F. A. (2018). Rate of recalibration to changing affordances for squeezing through doorways reveals the role of feedback. *Experimental Brain Research, 236*, 1699–1711.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Hacques, G., Komar, J., & Seifert, L. (2021). Learning and transfer of perceptual-motor skill: Relationship with gaze and behavioral exploration. *Attention, Perception, & Psychophysics, 83*(5), 2303–2319.

Hospodar, C., Franchak, J. M., & Adolph, K. E. (2023). Performance variability and affordance perception: Practice effects on perceptual judgments for walking versus throwing. *Experimental Brain Research, 241*, 2045–2056.

Ishak, S., Franchak, J. M., & Adolph, K. E. (2014). Perception-action development from infants to adults: Perceiving affordances for reaching through openings. *Journal of Experimental Child Psychology, 117*, 92–105.

Karasik, L. B., Tamis-LeMonda, C. S., & Adolph, K. E. (2016). Decisions at the brink: Locomotor experience affects infants' use of social information on an adjustable drop-off. *Frontiers in Psychology, 7*, 797.

Kingdom, F. A. A., & Prins, N. (2010). *Psychophysics: A practical introduction*. Academic Press.

Konczak, J., Meeuwsen, H. J., & Cress, M. E. (1992). Changing affordances in stair climbing: The perception of maximum climbability in young and older adults. *Journal of Experimental Psychology: Human Perception and Performance, 18*(3), 691–697.

Kretch, K. S., & Adolph, K. E. (2013a). Cliff or step? Posture-specific learning at the edge of a drop-off. *Child Development, 84*, 226–240.

Kretch, K. S., & Adolph, K. E. (2013b). No bridge too high: Infants decide whether to cross based on the probability of falling not the severity of the potential fall. *Developmental Science, 16*, 336–351.

Labinger, E., Monson, J. R., & Franchak, J. M. (2018). Effectiveness of adults' spontaneous exploration while perceiving affordances for squeezing through doorways. *PLoS One, 13*(e0209298).

Levy, I., Snell, J., Nelson, A. J., Rustichini, A., & Glimcher, P. W. (2010). Neural representation of subjective value under risk and ambiguity. *Journal of Neurophysiology, 103*, 1036–1047.

Linares, D., & López-Moliner, J. (2016). quickpsy: An R package to fit psychometric functions for multiple groups. *The R Journal, 8*(1), 122–131.

O'Neal, E. E., Plumert, J. M., McClure, L. A., & Schwebel, D. C. (2016). The role of body mass index in child pedestrian injury risk. *Accident Analysis and Prevention, 90*, 29–35.

Plumert, J. M. (1995). Relations between children's overestimation of their physical abilities and accident proneness. *Developmental Psychology, 31*, 866–876.

Plumert, J. M., & Kearney, J. K. (2014). How do children perceive and act on dynamic affordances in crossing traffic-filled roads? *Child Development Perspectives, 8*, 207–212.

Plumert, J. M., Kearney, J. K., & Cremer, J. F. (2007). Children's road crossing: A window into perceptual-motor development. *Current Directions in Psychological Science, 16*, 255–258.

Rachwani, J., Herzberg, O., Kaplan, B. E., Comalli, D. M., O'Grady, S., & Adolph, K. E. (2022). Flexibility in aciton: Development of locomotion under overhead barriers. *Developmental Psychology, 58*, 807–820.

Snapp-Childs, W., & Bingham, G. P. (2009). The affordance of barrier crossing in young children exhibits dynamic, not geometric, similarity. *Experimental Brain Research, 198*, 527–533.

Stefanucci, J. K., & Geuss, M. N. (2010). Duck! Scaling the height of a horizontal barrier to body height. *Attention, Perception, Psychophysics, 72*, 1338–1349.

Strother, L., & Kubovy, M. (2006). On the surprising salience of curvature in grouping by proximity. *Journal of Experimental Psychology: Human Perception and Performance, 32*(2), 226–234.

Tamis-LeMonda, C. S., Adolph, K. E., Lobo, S. A., Karasik, L. B., Dimitropoulou, K. A., & Ishak, S. (2008). When infants take mothers' advice: 18-month-olds integrate perceptual and social information to guide motor action. *Developmental Psychology, 44*, 734–746.

van der Meer, A. L. H. (1997). Visual guidance of passing under a barrier. *Early Development and Parenting, 6*, 149–157.

Wagman, J. B., & Malek, E. A. (2007). Perception of whether an object can be carried through an aperture depends on anticipated speed. *Experimental Brain Research, 54*(1), 54–61.

Warren, W. H. (1984). Perceiving affordances: Visual guidance of stair climbing. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 683–703.

Warren, W. H., & Whang, S. (1987). Visual guidance of walking through apertures: Body-scaled information for affordances. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 371–383.

Wichmann, F. A., & Hill, N. J. (2001a). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics, 63*, 1293–1313.

Wichmann, F. A., & Hill, N. J. (2001b). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & Psychophysics, 63*, 1314–1329.

Wilmut, K., & Barnett, A. L. (2011). Locomotor behavior of children while navigating through apertures. *Experimental Brain Research, 210*, 185–194.

Yasuda, M., Wagman, J. B., & HIguchi, T.. (2014). Can perception of aperture passability be improved immediately after practice in actual passage? Dissociation between walking and wheelchair use. *Experimental Brain Research, 232*, 753–764.