An Excel-Based, Curve-Fitting Method

For Developing Norms

by Christian P. Gruber

and the WPS R&D Staff

January 2013 Edition

# Contents

Chapter 1—Introduction. Domain of application, definitions and goals.

Chapter 2—Laying out the raw data. Setting up a spread sheet for the analyses. Coming to grips with the data and with the problems that the methods address.

Chapter 3—Normalized Estimation of Age-Group Medians and SDs. Going back a step. Looking at frequency tables for the raw data that lies behind means/SDs (linear approach) and medians/percentiles (normalized approach).

Chapter 4— Fitting Polynomial Trendlines to the Data. Smoothing the calculated values.

Chapter 5— Using Polynomial Equations to Estimate Smoothed Normative Means and SDs. Reviewing and revising the number of age groups to produce a more uniform gradient;

Chapter 6— Resolving Large Mean Differences Between Adjacent Age Groups. Re-expanding the number of age groups

Chapter 7— Checking Norm Performance in Critical Regions. Using final mean and SDs to calculate extreme standard scores and review for clinically critical regions

# Chapter 1: Introduction

This handbook describes a process for developing norms for tests of ability—that is, tests that yield an ordered upward progression of scores with age. To measure individual differences in ability, a child’s performance must be compared to that of same-age peers. While there are alternatives—“continuous norms” as proposed by former WPS Project Director Gayle Roid and statistical consultant Richard Gorsuch in the late 80’s—the standard approach in test development is to stratify the normative sample into age-ordered groups and calculate scaled scores (standard scores, t-scores, z-scores, etc.) within each group.

Question #1 is, “So, how many groups do we need?” In part, the answer depends on the characteristics of the test and the ability being measured. Some skills develop quickly—for example, most children master the formation of speech sounds (e.g., Arizona-3) by age six. Other skills develop more slowly—for example, most of the skills underlying “fluid intelligence” (e.g., Shipley-2 Abstraction or Block Design) don’t reach a flat or slowly increasing gradient until somewhere around 14 to 16 years of age, whereas a “crystallized” skill like Shipley-2 Vocabulary can show continuing growth well into maturity.

Furthermore, a test may have questions or items that intentionally divide the range of abilities into gross regions (as in the Shipley where just a few items cover a very wide span of abilities) or into thin slices (as in speech tests like the OWLS or CASL). Without careful item development, a test may have regions of sparse measurement (large gaps in difficulty between consecutive items), or regions with redundant items clustered in a narrow ability range (as in the original Shipley or Reynell).

So the answer to question #1 may emerge through one of several methods. With an entirely new test, it may involve an iterative, trial and error process where the characteristics of the ability are teased out by collecting samples, tweaking items, and repeating the process. For a revision, in contrast, there may be a prior edition that provides a clear template for stratification. For other projects, a more sophisticated IRT-analysis of pilot data may be used to identify item discrimination curves and the ability gradient.

No matter the approach, however, the process needs a criterion and that boils down to deciding “How big a change in skills are we willing to tolerate between adjacent norm groups?” I’ll jump in with both feet here and say “ONE THIRD OF A STANDARD DEVIATION.” I see three rationales for this:

1. During the past 20 years, the “1/3 SD” rule-of-thumb has gained broad acceptance in the community of test developers and publishers. A wide variety of intelligence, aptitude and achievement tests show the “Flynn Effect” —people’s ability, measured in raw score units, seems to drift upward about one-third of a standard deviation every ten years. This has generated another rule-of-thumb: ability tests need to be re-normed about every ten years. That is, test users can tolerate a systematic distortion of up to (but not more than) about one- third of a standard deviation, before they begin to lose confidence in clinical inferences based on test scores.
2. Another rationale is that one-third of a standard deviation seems to be a fairly natural “just noticeable difference” in clinical practice. So, in thinking about test results for a given child, clinicians tend to agree that a difference of 5 points on an IQ-type standard score (M = 100, SD = 15), or 3 points on a *t*-score (M = 50, SD =10), is right on the edge of warranting interpretation.
3. Finally, a third of a standard deviation is a bit larger than a “small effect size” but smaller than a “medium effect size”, according to the conceptual framework developed by Cohen (1988, 1992). Cohen’s work provides an index of the “meaningfulness” of group mean differences, and, analogously, a good metric for evaluating the clinical significance of differences between test scores.

If you look at norms tables for high-stakes intelligence tests (e.g., WISC, Stanford-Binet, etc.), you see that one-third of an SD jump holds up as the largest tolerated difference between consecutive norm groups, at least in the middle of the test’s age range. The rule-of-thumb has been applied less consistently to the younger age groups (larger ability jumps between adjacent age groups), or at the upper end (smaller differences between age groups). My guess is that these tolerated discrepancies reflect “cosmetic” decisions—test developers are reluctant to fragment the sample into multiple, tiny age segments on one end and are equally reluctant to have older children compared against others across what may seem like an exceedingly wide span of years. So, despite some idiosyncratic practices, it still makes sense to view “1/3 SD” as a benchmark.

Question #2 is, “So how big do these age-stratified groups need to be?” Broadly speaking, an N of 30 is about where SDs begin to stabilize. With test norms, however, it is sometimes desirable to have bigger groups, as it may require 50 or 100 individuals to assure that each group has an adequate sampling of demographic factors (e.g., gender, certain ethnic or socio-economic sub-groups, etc.). So the target sample size may be 30 to 100 in each age stratum, with the final number determined by the need for demographic representativeness. Larger age groups are advisable for high-stakes achievement or IQ tests, where gender, ethnicity, and SES may influence scores, and where results may be challenged on legal or other grounds. Smaller age groups may be sufficient for clinical tests whose primary purpose is to inform treatment design, or where the targeted ability is less affected by demographic variables.

So, the perfect set of test norms has equal numbers of children in groups of 50 or so with raw score means spaced exactly one-third of a standard deviation apart. Trust me, this never happens! The rest of this document is an exposition of how the best-laid sampling plans fall apart in reality, and the techniques that are used to reassemble the resulting data set into a well-functioning set of norms.

Please note that the following sections refer to Excel worksheets (e.g., Worksheet A), which are contained in the same folder as this handbook. When you open these sheets to work with them, please immediately *Save As* to another location, to preserve the original worksheets for the next reader.

# Chapter 2—Laying Out the Raw Data

**Worksheet A** shows the raw score means and SDs for the Reading Comprehension test from the OWLS-2 (with thanks to Amber Klein for overseeing the data collection and developing the coding and scoring rules, to Dave Herzberg for developing the programs for scoring the data set and assembling the raw scores, and to consultant Colin Elliot for illuminating the basic principles of this norming method). The OWLS-2 provided the springboard for developing these current norming practices.

The *norm grp* column identifies the starting age in years and months for each of the norm groups (50 is 5 years 0 month, 59 is 5 years 9 months, etc.). Note that these--the original norm groups built into the data collection--were at 3 month intervals for the first two years, then 6 month intervals, then 1 year and finally multi-year intervals (e.g., group 160 covers 16, 17 and 18 years, while 190 covers 19, 20 and 21 years). These were the *a priori* age breakdowns used during data collection, which resulted in fairly uniform samples sizes of around 30-50 individuals per group.

The *mean* and *SD* columns show the raw score means and standard deviations of these groups, while the *EFFECT* column shows the standardized difference between adjacent means. This latter value, otherwise known as the *effect size* (Cohen, 1992), is simply the raw score difference between two means, divided by their pooled (averaged) standard deviation. It is the metric we use to evaluate how our age groups fare against the “1/3 SD” rule-of-thumb.

At first glance, the *a priori* groups seem to work well: the average effect size for the differences between adjacent groups was exactly .33. Yet the individual effect sizes vary quite a bit, and, consequently, many are larger than is desirable. Indeed, 6 of 22 have an effect size of .45 or greater, and one at age 12 years shows a jump of .88, almost a full standard deviation. And offsetting these are groups that differ little from their immediate predecessors (e.g., age 9 years and 13 years) or, at age 19, violate the developmental expectation and appear to show students losing skills as they grow older.

Closer examination reveals two types of problems: A general issue with groups that are spaced unevenly along the curve of typical development; and local issues reflecting gross fluctuations and reversals of the expected progression of scores. It is tempting to look at the local problems and work out a local solution. For example, note that the big jump at age 12 is followed by the non-jump at age 13, suggesting that the “average” progression is smoother and that the problem may lie in the unique group of 54 kids in the age 12 sample. You may note that the SD of this group is lower than adjacent groups. So perhaps we might look at this group in detail and see if there are bad data points? It turns out to be better to wait, however, and develop a fuller review of the data before working on solutions.

As a second step in understanding the raw data, we can graph the means against a simple ordinal scale (that is, 1, 2, 3, etc.; shown in the *ordinal grp* column) with an implied equal distance from group to group. This makes sense in that we have aimed to have groups fall on equally spaced intervals, so we review them against such a template*.*

*First, drag/highlight the “mean” column, including the header as well as the numerical values. Then go to the Excel “Insert” tab, find the “Charts” option and within that the “Scatterplot” option. Select the simple scatterplot with no connecting lines*. *To get rid of a graph, click on it to select it, then hit the delete key.*

The resulting graph has been included in Worksheet A, but go ahead and recreate the graph yourself for practice.

The problem with the jump between age years 12 and 13 (ordinal groups 17 and 18) is clear in the graph. And one “brute force” fix seems obvious—just average the gap to adjacent means in both directions for the age 12 group. This implies the use of adjacent cases in the same region, as a way of increasing our sample size at the problem point and getting a better estimate of the underlying mean. In a sense, the method of this handbook follows the same process, but in a much more systematic and useful way.

Before proceeding, however, there is another important message in the raw data. The SD values are not stable. In the middle of the age range, the SD varies quite a bit, from as low as 12.4 at age 7 years 6 months to as much as 26.1 at age 15 years. Even more problematic, the variation appears to be systematic. Particularly at the young ages, but also among the teenagers, SDs tend to be smaller than those seen through the middle of the tested ability range. Graphing the SDs can help make the systematic nature of the problem more evident. Follow the same instructions provided earlier for graphing the means and apply it to the SD column. You will see in the resulting graph that the SD distribution forms an inverted “U” shape, with low values in the youngest groups, rising to a peak at the 15th ordinal group and then falling off again, particularly in the oldest four groups.

This suggests that the test may produce both floor effects (many young kids with low scores pulling the mean toward zero and compressing the SD) and ceiling effects (many older kids with near-perfect scores pushing the mean toward the top of the score range, and compressing the SD). When these effects operate, the raw SDs may yield a distorted picture of a particular child’s position relative to the mean of his or her norm group. Within the youngest age groups, children who score above the mean of their same-age peers will appear more talented than is actually the case. Analogously, within the oldest age groups, children who score below the mean of their same-age peers will appear more impaired than is actually the case. These floor and ceiling effects are demonstrated clearly in the next chapter.

# Chapter 3—Normalized Estimation of Age-Group Medians and SDs

**Worksheet B** contains frequency tables (copy-and-pasted from SPSS output) for each of the original Reading Comprehension norm groups. These are the same groups whose means and SDs were shown in Worksheet A. The frequency counts reveal skewed score distributions in many groups. In the table for age 5 years 0 months, for example, nearly 9% of children got zero scores, more than 40% got three or fewer answers right and 65% got 5 or fewer right (the rightmost column shows the cumulative percentage).

When scaled scores (e.g., T-scores, standard scores, z-scores, etc.) are calculated directly from means and SDs, they are called “linear” scaled scores. Deriving scores in this way with the youngest norm group (M = 6.3, per Worksheet A) would result in about 70% of the normative sample being labeled “below average.” Clearly, this won’t work.

An alternative is to derive scaled scores from the actual score distributions in each norm group (instead of the distribution “moments”, e.g., means and SDs). In this method, we use the cumulative percentage associated with each raw score to assign its standard score. For example, 84% of scores are at or below the first standard deviation above the mean, so a raw score associated with a cumulative percentage of 84 would be assigned a standard score of 115 (or a T-score of 60, or a z score of +1). Scores calculated in this way are called “normalized” scaled scores. When raw scores within normative groups are normally distributed, the linear and normalized approaches yield nearly identical raw-to-standard-score conversion tables. (In this context “nearly identical” means “functionally equivalent for psychological and educational measurement”.)

To see how this equivalence works in “well-behaved” data, skip down to the table for kids 9 years 6 months in age. Looking down the right-most column (cumulative percentiles) you will see that 52.8% of children got 63 or fewer correct answers, so the 50%ile is imputed to be slightly below 63. This value is nearly equivalent to the mean (62.24) reported for this group on Worksheet A. Farther down the cumulative distribution column in Worksheet B, 86.1% of children got a raw score of 87 or less, so a child scoring at the 84th %ile (corresponding to +1 SD), would likely show a raw score in the range of 83 to 87. Referring to Worksheet A, the raw score point for the mean plus one SD is 62.24 + 21.31 = 83.5. Again, in a norm group where the raw score distribution approximates normality, the linear and normalizing approaches produce comparable results.

But we already know that in the OWLS-2 Reading Comprehension data, many of the original norm groups have skewed, non-normal raw score distributions. How do we get from these actual frequencies and percentages to the “means” and “standard deviations” needed to calculate scaled scores? Even though the normalizing approach is not strictly needed for the 9 years, 6 months group (which starts on Worksheet B at row 319), we’ll use this group to illuminate the process.

On Worksheet B, to the right of the frequency tables are a series of numbers. For the 9 years 6 month group, two blocks of numbers have been highlighted. In the larger block, note that the number 50.00 is adjacent to the first cumulative percentage greater than 50. To the right of the 50.00 is 62.50, which is the estimated raw score at the 50th %ile. If you click on the cell with 62.50, you will see the underlying calculation in the formula bar. It shows how the 50th %ile score has been imputed from the frequencies in the table. We now have a value (62.50) for the *median* score, one that we will use in our norm calculations *as if it were the mean*.

Now look at the smaller highlighted block at the top of the age 9 years 6 months distribution, (Worksheet B row 321). You will see that the number 5.00 is adjacent to the first cumulative percentage greater than (or equal to) 5. To the right of the 5.00 is 30.8. If you click on this cell you will see the underlying calculation. It shows how the 5th %ile score has been imputed from the frequencies in the table. The same logic applies to the calculation of the 10th %ile just below, and the 90th %ile and 95th %ile scores at the bottom of the 9-6 table.

Go back to the larger highlighted block starting at row 332. Click on the cell with 31.70, which shows the raw score difference between the 50th %ile score and the 5th %ile score. Now click on the cell to the right of that, showing 19.27. In this cell, the difference between the raw scores at the 5th and 50th %iles has been divided by 1.6449. This denominator is used because in a normal distribution, the 5th %ile raw score lies 1.6449 SDs from the mean. Thus, using a calculation based on the 5th %ile raw score, we have estimated the SD of this norm group to be 19.27. In the same columns of the table, in nearby cells, you will find that the SD has been estimated four times, corresponding to the raw scores at the 5th %ile, 10th %ile, 90th %ile and 95th %ile. There are thus two pairs of SD estimates, one above the median, and one below the median.

In the next column to the right, we have averaged each pair of estimates, resulting in an estimated SD of 19.52 for children scoring below the mean and 23.85 for those scoring above the mean. For subsequent analyses we will continue to carry forward separate SD estimates for the upper and lower halves of the distribution. But to complete this opening “proof,” it is worth noting that the average of 19.52 and 23.85 is 21.68 and that the conventionally calculated SD in Worksheet A is 21.31. Close enough for test development purposes.

To finish this process, scroll back up to the first frequency table, the one for the children aged 5 years 0 months (starting at Worksheet B row 6). The 5-0 table demonstrates the normalizing approach, using the 50th %ile to find the median score of 3.6 (which now serves as an estimate of the “mean”), then the 90th and 95th %iles to estimate an “SD” of 5.36 for the children scoring above the median, and the 5th and 10th %iles to estimate an “SD” of 2.83 for children scoring below the median. The example illustrates two central features of the normalizing method: (1) the use of cumulative percentages to identify “means” and “SDs”, and (2) the development of two “SD” estimates per norm group, one above and the other below the “mean”.

The advantage of separate SD estimates is easily apparent within the 5-0 normative group. Note that in the higher performing half of the sample, the SD estimate (5.36) is about twice the size of that for the lower performing half (2.83). If we were to evaluate higher performing children against the smaller SD of the lower half of the distribution, they would be credited with much more extreme apparent performance (standard scores) than their percentile placement would warrant. And the opposite would be true of the children in the lower performing half—their apparent performance (evaluated against the SD of the higher performing children) would be much less extreme than would be warranted by their percentile placement. If we used a single SD estimate (the average of the high and the low, or the linear SD) these aberrations would be less extreme, but nevertheless a serious distortion of desired interpretive meanings associated with standard scores

On the next table (5 years 3 months), the difference between the upper SD (12.59) and lower SD (3.86) is even larger and more discrepant from that taken from linear calculations (7.4 from Worksheet A). At these young ages, the advantages of using the normalizing approach and split SD estimates are clear: many more children will be properly placed close to their actual percentile ranks.

Also check the upper-age frequencies. In the age 19 years and up table, for example, the lower estimated SD is 22.84 while the upper one is 7.65, suggesting that higher skilled kids are bunching up against the ceiling of the test. In this range, a linear calculation SD estimate (19.65 from Worksheet A) would mischaracterize the abilities of the best students.

These analyses show that the normalizing approach will be necessary for at least some groups at both extremes of the tested age range. It is also clear that the two approaches will produce similar results for groups in the middle of the ability range, where there has been no compression of the SD. For these reasons, we will just use the normalizing approach and carry forward all three sets of results (medians to represent means and separate estimates of lower SDs and upper SDs) for all norm groups into the next chapter.

### NOTES

You can copy and paste the highlighted blocks of calculations from one frequency table to the next, but note that percentile blocks must each be pasted separately as they each contain a hard coded number that is different for each estimated percentile. For example, the block for the 90th %ile in 5-0 table (row 33) can be copied to the appropriate line in the 5-3 table (row 57) and so on down through all of the tables for the different groups. The formula within that block contains the number 90 to represent the 90th %ile. But the 95th %ile block must be copied separately, because the formula within the cell contains the number 95, instead of 90.

Also, once you paste the large block containing the 50th %ile calculation and all the SD estimates, you will have to go in by hand and supply the cell references for the formulas. These will differ from table to table, depending on the score distributions. It’s tedious, but it is really the only substantial “rote piece” of this approach.

There may be cases where more than 5% of children get zero scores or when the highest achieved score is obtained by more than 5% of students. In these cases, the logic presented earlier and reference to a table of normal curve percentages (from a statistics textbook or the internet) will enable you to estimate an SD from other percentage points. Try to use fairly extreme points, however, as the systematic distortions will be most evident near the extremes of the distributions.

I chose to use OWLS-2 Reading Comprehension for this Handbook because it shows both floor and ceiling effects. It’s probably more common for a given test to show one of these effects and not the other, or to show either effect across a smaller set of norm groups. In those cases, you can use the normalizing approach on only a few groups, and retain the linear means and SDs for the other groups. When combining the normalizing and linear approaches, it’s good practice to do quite a few of these frequency table analyses, beyond the point where linear estimated SDs “look”OK. If, from table to table, there continue to be systematic differences between the upper and lower SDs, it’s better to use the normalized values.

I chose to use the five non-parametric (percentile) points to estimate the required statistics and, in subsequent smoothing, actually smoothed only three of them. Colin Elliot seemed to think this sounded sensible. I have seen a paper by Gayle Roid where it appears that quite a few more points on the frequency table were used to estimate SDs and then to have been smoothed separately. In addition we have heard that a PsychCorp statistician developed programs that smoothed based on formal parametric statistical moments (i.e., skewness and kurtosis in addition to mean and SD). My guess is that these approaches are automated—that is, using programming within SPSS or Excel, to circumvent the need for direct inspection of the data for individual norm groups. Based on my experience developing the current method, I doubt that an automated approach would save much time or produce “better” norms. Seeking shortcuts to avoid the admittedly tedious “hands on” inspection or group-by-group percentile distributions may lead to unwelcome “surprises” down the road. This is my intuition, and I’m happy to be proven wrong by someone with the time and inclination to do a thorough review/demonstration of an “automated” process.

# Chapter 4— Fitting Polynomial Trendlines to the Data

**Worksheet C** brings forward the values we calculated in the Worksheet B frequency tables for each of the a priori normative groups--a median and two separate estimates for the SD, one for the scores in the lower-performing half of the norm group and one for the scores in the higher-performing half of each age group (Worksheet C lines 3 to 28). We are a step closer in having SD values that will place children more accurately. Yet there are still anomalies in the differences between adjacent group medians--the average effect size at .34 is about the same as in Worksheet A, but there are two reversals and two effect sizes greater than .70.

For many tests we might start estimating trend lines at this point. The OWLS-2 research project, however, included a parallel Form B and accompanying data set. Needless to say, the progression of age group means and SDs did not march in lock step between the two forms! Decisions about how to stratify the final norm groups were therefore constrained by the need to maintain parallel structure between Forms A and B. We also made some decisions based on a general need to preserve usability and clarity in the final test materials. For example, once the developmental curve flattened enough to support six-month (as opposed to three month) age intervals, we did not go back to three-month intervals to break up a big jump at the older ages.

If you scroll down to the lines below the first set of figures in Worksheet C (lines 32 to 50), you will see the final outcome of several iterative sets of such review and decision making. Comparing the upper and lower versions, you will see, for example, that at 8 years , 9 years and 10 years, the six-month groups have been combined to make the steps at one-year intervals. To this stratification scheme, we will now apply the first stages of the smoothing process. Notice that we now have 17 rather than 22 norm groups and that these are numbered consecutively in a column to the left of the data. The same basic problems are still present: large and uneven jumps from group to group.

The first task is to establish a polynomial trend line for the median data. Start by graphing a scatterplot of the medians, following the instructions provided earlier in graphing the means.

* *Then select (click on) the graph of the scatterplot for the medians. The Chart Tools tab will light up. Within that tab select Layout, within that select the tab for Trendline, and within that select Exponential. A line will appear running through the data. Double click on the trend line itself. A box will open up. Select Polynomial. The line will change to a second order polynomial (single arc or bend). The box will stay open. Now, within the polynomial choice, change the “2” to a “3” to see the fit of a third order polynomial, one with two bends and an S shape. You will see that this third order polynomial comes closer to more data points.*

To provide some context, second and third order polynomials fit the estimation of means for virtually all data I have worked on and curves of these kinds generally have a good rational: a gradually diminishing developmental growth curve is a simple arc (or second order polynomial) that captures a process where a skill is learned in larger chunks and then refined with smaller changes approaching maturity. The “S” shape (or a third order polynomial), on the other hand, often fits when a test covers a very broad age span, where the skill emerges slowly over the first few age groups, then steeply as it is acquired and integrated, then slowly again as complete maturity is reached.

Follow the procedure outlined above and fit curves to the values in the LO and HI SD columns. For this Reading Comprehension data it appears that a third order seems to fit the LO SDs, and a second order curve fits the HI SDs. Even given those systematic trends, it is clear that both sets of SDs also vary widely between adjacent groups. There is nothing very systematic in these larger fluctuations from group to group, however and so we assume that it is due to simple random effects in the small subsample sizes. For this reason, we do not worry about trying to create a close fit for a line, just one that roughly captures whatever systematic change is observed.

In fact, in the absence of any clear curvilinear trend it may be acceptable to use a single SD estimate across all groups (an implied straight horizontal trend line). In any case, a “well-smoothed” set of SD estimates will ensure that norms perform properly in the extreme regions (+/- 2 SDs), where clinical interpretations are most often made.

# Chapter 5—Using Polynomial Equations to Estimate Smoothed Normative Means and SDs

The next step is to impose these smoothed trend lines back onto the data set. **Worksheet D** brings forward the same tables of means and SDs and the associated graphs shown in Worksheet C. The first new element is that each graph now shows an algebraic formula. *To see how these are created, go to the medians graph and open the trend line box by double clicking on the trend line itself. At the bottom of the dialog box, find the line for “Display Equation on Chart” and notice that it is checked*. The equation is a standard algebraic expression for a curved line. As a third order polynomial it has separate elements for cubed, squared, and simple terms as well as one for the intercept*.* Doing the same for HI SD graph will show the equation for a second order polynomial—that is, lacking the cubed element.

The formula in the graph is first transferred to the Excel sheet. In Worksheet D the work has been done as you will see in the parallel table to the right of the first table. It is best to set the table up by putting in the column head and, most importantly, to have a column somewhere to the left where the ordinal integer labels are provided for the norm groups. In Worksheet D it is the first column on the original table, numbering from 1 to 17.

The first formula will be copied into the first cell of the new table, immediately under the Median heading. If you put the cursor on that cell in the prepared table, you will see the Excel formula that corresponds to the polynomial formula printed across the graph for Medians.

* *To make this transfer, go to the graph for the Median, click and drag across the formula to highlight it, and Copy it into memory (the clipboard). Then move the cursor onto the Excel file and click on the cell where the first mean will appear, DON’T PASTE YET. With the target cell highlighted, go to the top of the Excel sheet and find the formula bar where functions are typed in: now Paste the formula into that box.*
* *Then modify the formula to the proper Excel format seen in the example in Worksheet D. Edit the formula so that it starts with the equal sign (i.e., remove the “Y” and any leading blanks). Replace the “x” with the cell identifier that corresponds to the norm group integer ID for the “x” value. Add the “^” character to encode the power. When all edits are in place, taking the cursor out of the box (first place it at the end of the formula in the box, then place it in any unused cell in the table) will result in the display of a number: the mean for that group as it has been calculated by that formula.*
* *Now it gets easier. Simply “pull down” the formula that you edited, so that it is duplicated in the cells below for the other norm groups: Highlight the box now containing the correct formula and drag the lower right corner down to fill the cells below with the formula (the cell references will adjust automatically to refer the proper norm group ID for each row.)*

Follow the same processes to fill in the cells for each of the SD columns: Copy the formula from the graph to the Excel cell for the first norm group. Modify the formula to fit Excel conventions. Then drag the corrected formula from the top cell down to fill in the other cells for the column.

Now review the data. With the means and SDs in place you can drag down the column for the effect size. You will see substantial progress: uniformly increasing mean scores with very similar effect sizes between adjacent groups. The only problem is that the jumps, while uniform, are now all around .50, (1/2 SD), considerably larger than the ideal of 1/3 SD.

### Notes

I have tried to keep things “simple” in writing his handbook. With a real set of published norms at stake, the process would be more extended and would probably go back through several loops, smoothing and then looking at curves and considering combining groups. Indeed, that describes the process for the OWLS-2 Reading Comprehension norms.

It is also common and acceptable practice to “hand smooth” – that is, to choose sensible values for means and SDs for the youngest and oldest age groups. The need for hand smoothing arises because these extreme age groups do not have data points on both sides to help shape and constrain their values. Then too, as values that may also occur at the very end of a test’s discriminating age, their “typical values” may be poorly estimated, particularly in cases where there are strong ceiling or floor effects. In these cases, rather than attempting to create a unique calculation, it is usually better to examine nearby cells from the table of smoothed values and then those from parallel cells in the prior table of un-smoothed values and choosing final values that seem sensible and consistent with the developmental trends embodied in the rest of the age groups. Think through how kids would be placed (i.e., the standard scores they would receive) for the means and/or SDs you choose. For the groups at the extremes of the age range, values taken from this thought process are more likely to produce good norms than those taken automatically from smoothing formulas.

# Chapter 6—Resolving Large Mean Differences Between Adjacent Age Groups

**Worksheet E** brings forward the smoothed medians and SDs that were produced on Worksheet D. Now scroll down a panel and find the same table “expanded.” Using the “effect size” column for a guide, you will see that we have inserted a new row between most of the rows where median/means jumped by more than our target third of a standard deviation. As in other parts of the process, there were some rational constraints. For example, we inserted rows consistently so that the number of (now implied) age groups gradually and consistently diminished: four groups per year (3 month intervals) ages 5 through 7, two per year (6 month intervals) ages 8 to 14, then one per year and finally multiple years in one for ages 16 and up. Then we filled in the integer column for the new lines: 5.5 added between 5 and 6, 6.5 added between 6 and 7, and so forth.

Now go down to the third panel. You will find the same expanded table, but now with the missing cells filled in for the other columns. We did this by “pulling down” the formulas that were contained in the cells that had been carried forward from Worksheet D. For example, the median cell for the new “5.5” row came from pulling down the formula in the cell above it. Actually, it is easiest to do all the cells at once: just highlight the median, effect size and two SD cells for the top row (5 group, age 600) and pull the four columns down for the rest of the table. The formulas generate correct values for existing and new rows. We have now applied the smoothing polynomial equations to fill in the new age intervals.

This procedure moves us much closer to optimal norms: the effect sizes for the comparison between adjacent group means were halved and, in all cases where the subgroups were added, are now less than a third of a standard deviation.

Despite these improvements, larger jumps remain for 5 years 9 months over 5 years 6 months (.44 SD) and 6 years 0 months over 5 years 9 months (.52 SD). Using the same techniques it would be possible to estimate intervening values. A number of practical issues arise—for example, the need to also create smaller increments in the early half of age 5 and to choose a sensible interval (monthly norms?). And there are data driven concerns--the actual underlying N’s were already quite small and I felt uncomfortable stretching them over more discrete intervals; the actual raw scores in these ages were quite low and significant number of kids were already getting zero scores, no matter how you look at it. So, in this case, the decision was made to live with the larger intervals, subject to checking for consistent interpretation as described in the next chapter.

There is also an undesirably large jump in the means for the last group. Given more time (and a real rather than pedagogical purpose), I would probably have gone back and seen whether breaking these large blocks of kids into more groups, prior to smoothing, might make for a useful improvement.

As a final step, look down at a fourth panel. All of the columns of figures in the table actually represent the product of calculated formulas. It is time to change them to simple digits, in preparation to passing them on to either other staff (IT, RA) who will use them to calculate complete norm tables for the different applications or products. *Simply “Copy” the calculated table and “Paste Options” using “Values” rather than functions.* I also took the opportunity to apply less cryptic labels to the norm groups.

# Chapter 7—Checking Norm Performance in Critical Regions.

**Worksheet F** brings forward the smoothed and filled medians and SDs that were produced on Worksheet E. To the right of the table, you will find six new calculated columns (the median column is simply a copy and placed for convenient reference). Place your cursor in the fourth cell of the “-1 SD” column, where the number “4.77” appears. You will see that it represents the raw score value associated with a scaled score one SD below the mean—the mean is 10.7, so a score one SD below the mean is 10.74 minus 5.97 or 4.77. The “test” of well performing norms at this stage is that all scores within a column show uniformly rising values. In the present case, going to the next norm group (the “8.84”) demonstrates that a higher raw score will be required to get the same scaled score.

In the – 2 SD column, the same logic is applied to calculate raw scores that are two SDs away from the mean. Looking down the -2 SD column, you will see that the values always increase from row to row. The left-most column “Raw Score Diff” calculates the difference between adjacent means. It provides a visually salient check as a negative sign represents a “flip” in the table, a place where a child who gets a give low raw score would get a higher scale score when aging up – clearly an undesirable outcome.

On a diagnostic test like the OWLS, – 2SD is a critical region, the area where children are going to be seen as having significant problems. It is important to check at least this far out from the mean, because differences in SDs from group to group are amplified at more extreme values. The check performed in this case assures that scores as low as a standard score of 70 will be well behaved. If more extreme scores will be provided in norms tables, it is best to check them as shown here. It is time-consuming and potentially error-prone to calculate them out and then, after the fact, review the tables for reversals.

The columns to the right show the same kinds of checks run for high scores. The performance of norm tables in the higher regions are most critical when a test is used to identify talent—for example, to qualify for gifted programs, admission to highly competitive academic settings. The logic is the same and uniformly rising raw scores are required at the higher points on the distribution.

A few last quibble points. You will note that – 1 SD and -2 SD columns contain negative values for the young ages. This is the result of the floor effect for the test, the fact that so many children got zero and other very low scores, that for scale scores to exist below -1 SD, they would have to be below zero. Obviously impossible, so these kind of values must be blocked in norm calculations.

In general, there are problems with scale scores calculated on very low raw scores like 0, 1 and 2. Children performing this low seem scarcely to have established that they have a basic inkling of the tested content and assigning scale scores seems to be reifying something that is best expressed in cautious professional language rather than a “test score.” Expectations of users based on experience with earlier editions, or the practices seen in competitor instruments may overrule such scruples. In these cases, it’s wise to add some cautionary language when discussing interpretations.

If you look closely, you will also note a flaw at the positive extreme for the oldest students. At positive 2SD’s there is a “flip” and a given raw score might produce a higher scaled score when the student aged up. When these happen, it seems fine to “hand smooth” values that produced the flip, particularly since very small adjustments to a SD can produce an adequate fix at extreme values yet change the intervening scores very little. In this case, simply raising the “HI SD” value from 7.91 to 7.95 fixes the problem as far out as 2 SD. These isolated fixes may be a “slipperly slope”, however, with slight tweaks to fix one problem creating new ones. In the end, it may be better to resmooth the SDs with a differently polynomial curve, with the hope of producing more consistent values.