THE IMPACT OF EDUCATION AND EXPERIENCE ON DIAGNOSTIC ACCURACY

By

SUSAN I. GROSS

Submitted in partial fulfillment of the requirements for the degree of

Master of Arts

Department of Psychological Sciences

CASE WESTERN RESERVE UNIVERSITY

August, 2015
We hereby approve the thesis/dissertation of Susan I. Gross candidate for the degree of Master of Arts: Psychology*.

Committee Chair
Elizabeth J. Short

Committee Member
Robert L. Greene

Committee Member
Brooke N. Macnamara

Date of Defense
May 14, 2015

*We also certify that written approval has been obtained for any proprietary material contained therein.
Table of Contents

List of Tables ................................................................................................................................. 3
Abstract ........................................................................................................................................... 4
Introduction ....................................................................................................................................... 5
Methods ........................................................................................................................................... 11
  Raters ........................................................................................................................................... 11
  Materials ....................................................................................................................................... 12
    Play Videos ................................................................................................................................. 13
    Rating Form ............................................................................................................................... 15
Procedure ......................................................................................................................................... 16
Data Analysis .................................................................................................................................... 16
Results ............................................................................................................................................ 17
Discussion ....................................................................................................................................... 20
  Limitations ................................................................................................................................... 23
  Clinical Implications ..................................................................................................................... 26
Tables ............................................................................................................................................. 28
References ...................................................................................................................................... 32
List of Tables

Table 1: General diagnostic accuracy ................................................................. 28
Table 2: Specific diagnostic accuracy ................................................................. 29
Table 3: Rates of correct and incorrect classification ........................................ 30
Table 4: Misclassification and effect sizes ......................................................... 31
The Impact of Education and Experience on Diagnostic Accuracy

Abstract

by

SUSAN I. GROSS

This study was designed to examine the role that education and experience play in diagnostic accuracy. Four groups of raters, differing in education (low or high) and experience (low or high), were asked to categorize 93 children with typical development, specific language impairment (SLI), attention-deficit/hyperactivity disorder (ADHD), or autism spectrum disorder (ASD) based on their performance during a standardized five-minute free play assessment. After reviewing the child’s play, raters categorized the child into general developmental category (typical or atypical) and specific diagnostic category (typical, SLI, ADHD, ASD) and then ranked their confidence in each categorization. Education and experience did not contribute to accuracy generally or specifically. However, raters classified children with typical development more accurately than children with ASD, ADHD, or SLI, and they also had more success identifying ASD than ADHD or SLI. Overall, raters were more confident when the classification was correct than when it was incorrect.
The Impact of Education and Experience on Diagnostic Accuracy

Specific language impairment (SLI), attention-deficit/hyperactivity disorder (ADHD), and autism spectrum disorder (ASD) are developmental disorders that show promising evidence for improved outcomes with early intervention (Gillon, 2000; McGoe, Eckert, & Dupaul, 2002; Rogers, 1996). Before early intervention can take place, children must first be identified. While referral to appropriate professionals is known to be an important factor in determining when a child is identified (e.g., Mandell, Novak, & Zubritsky, 2005), less is known about the characteristics of the people referring children to services that may change their ability to identify worrisome behavior in preschool children. In order for an adult (i.e., parent, teacher, family friend) to recommend referral for diagnosis, he or she must have identified a behavior or characteristic in the child that appears outside of the range of typical development. Education level and background experience are assumed to affect diagnostic decision-making, so the current project examined whether these factors do in fact influence the ability to identify typical and atypical development in a diverse sample of preschoolers (i.e., SLI, ADHD, ASD, and typical development). Children are referred for evaluation services based on the observations and opinions of the adults they encounter, so the present study seeks to understand whether the education and experience of adults change their accuracy in identifying problematic behavior.

The preschool setting presents cognitive, attention, social, and emotional demands to young children that are not commonly encountered in the home-setting, including but not limited to demonstrating peer-appropriate behavior, staying on task, and following specific directions. These new demands often increase the likelihood that problematic
behaviors will emerge, be recognized as different from typical behavior, and be identified as long-term threats to optimal development (e.g., Spira & Fischel, 2005). Given new task demands, it is not surprising that rates of diagnosis increase for developmental disorders during the preschool period, including SLI, ADHD, and ASD (Hertz-Picciotto & Delwiche, 2009; Mandell et al., 2005; Tomblin, Records, Buckwalter, Zhang, & O’Brien, 1997).

Although children are more likely to be diagnosed with SLI, ADHD, or ASD in the preschool period than in infancy or toddlerhood (Hertz-Picciotto & Delwiche, 2009; Mandell et al., 2005; Tomblin et al., 1997), early identification still proves challenging for many reasons. First, differences in functioning can result simply as a natural byproduct of development, with some children acquiring skills at a faster rate than others. Individual differences in the timing of the appearance of skills are common, with the developmental windows for skill acquisition quite variable (e.g., single word acquisition can occur between 9 and 18 months; Bates, Bretherton, & Snyder, 1991). Ascertaining when the rate difference is diagnostic or simply a part of normal development is challenging. Thus, the line between typical and atypical development is very difficult to establish. Second, disagreement about developmental category can arise in part because some behaviors identified as problematic in the preschool period remit on their own or become less problematic as the child develops (e.g., Faraone, Biederman, & Mick, 2006). In spite of difficulties associated with the identification of developmental disorders in preschool children, diagnostic classification must occur in order for children to receive appropriate services during the optimal developmental window (Ramey & Ramey, 2004). As such, early identification has been of great interest to practitioners, teachers, and
parents alike. While both child and observer characteristics are important in the identification of developmental risk, the present study focused primarily on the impact of two important observer characteristics: education and experience.

It is often difficult to disentangle the effects of education from experience on diagnostic accuracy, particularly given the integrated nature of most training programs. Research on diagnostic expertise in psychology began over fifty years ago (Goldberg, 1959). Hypothesizing that both education and experience have an impact on diagnostic proficiency, Goldberg compared three groups of raters differing in training level (non psychologists with no noteworthy experience, M.A. psychology trainees with 1-4 years of experience, and Ph.D. psychologists with 4-10 years of experience) on their accuracy in differentiating patients with an organic impairment (i.e. trauma, tumor, thrombosis) from those with a non-organic impairment (i.e. schizophrenia, anxiety neurosis) based solely on the patient’s performance on the Bender Gestalt Test. None of the rater groups showed superior diagnostic performance, and therefore no effect for education or experience was demonstrated. Questions of ecological validity should be noted, as raters were not able to directly observe the patient. In addition, only the accuracy levels for the etiology of brain injury was examined; therefore, it is unclear if these effects generalize to other diagnostic domains such as developmental disorders.

In addition to a lack of effect in diagnostic accuracy for brain injury, training level also did not improve diagnostic accuracy between novices (clinical psychology master’s students with less than a year of experience) and experts (clinical psychologists with an average of 20 years of experience) reading case study descriptions of two adults, one with co-occurring panic disorder with agoraphobia and depressive disorder not otherwise
specified and the other with co-occurring borderline personality disorder and depressive disorder not otherwise specified (Aarts, Witteman, Souren, & Egger, 2012). In spite of equal diagnostic accuracy levels, experts were more confident in their diagnoses than novices. Additionally, research by Witteman and van den Bercken (2007) demonstrate that the relationship between experience and diagnostic accuracy may not be linear. Interestingly, novice clinicians with minimal experience and experts with more than 16 years were better able to diagnose child and adult cases of psychopathology based on vignettes than clinicians with an intermediate amount of experience (2 to 15 years). While it is not possible from the study to completely account for these findings, one explanation may be that the intermediate group had enough experience to cause them to question their diagnosis but not enough to home in on the most accurate one. The results may be limited to diagnosis based on verbal vignettes.

In contrast, evidence for the positive impact of training was shown in a unique paradigm in which participants determined diagnosis based on a simulated clinical interview via a computer (Brammer, 2002). Participants were asked to arrive at a diagnosis for a client based on a verbal description, with their decision-making further guided by the ability to ask questions and receive appropriate responses from a computer. Responses to questions indicated the client had symptoms of major depression, Bipolar II disorder, alcohol abuse, acute stress disorder, borderline personality disorder, or nicotine dependence. Path analysis, using education and experience as continuous variables rather than groups of novices and experts, revealed that education (bachelor’s degree, master’s degree, doctoral student, or doctoral degree) did not have a significant direct effect on accuracy, but experience (varying from 0 to 28 years) did. Education did have an indirect
effect, where increased education led to increased numbers of questions about patient and family history, which in turn was associated with stronger diagnostic accuracy. In spite of positive evidence for the influence of education and experience on diagnostic accuracy following a simulated clinical interview, the effect may have been due to differences in questions asked, reflecting differences in interview styles, rather than true differences in diagnostic accuracy. It is not clear whether education or experience would impact diagnostic accuracy levels if raters received consistent information. In addition, a computer gave interview information, so human characteristics such as intonation and appearance could not be used to formulate hypotheses.

It is alarming that most studies do not show an increase in diagnostic accuracy as a function of training, as our advanced educational system is founded on the premise that expertise is achievable through education and experiential learning in a particular domain. In regards to experience, expertise in general was first speculated to require 10 years of practice by Simon and Chase (1973) in their study of chess expertise. The ten-year-rule has been expanded to other areas such as music (Hayes, 1981) and sports (Monsaas, 1985) but has yet to be established in diagnostic realms such as psychological assessment. In a meta-analysis, deliberate practice, defined as specifically targeted, repetitive skill use with modifications based on repeated assessment, feedback, and increasing difficulty (Ericsson, Krampe, & Tesch-Römer, 1993) did account for some of the variance in performance in domains such as music, sports, and chess, but it did not show strong accounts for explaining variance among performance in professions or education (Macnamara, Hambrick, & Oswald, 2014). Therefore, it is unclear if education and experience provide the needed practice to improve diagnostic accuracy or if other
characteristics are more important. Dawes (1994) argues that psychologists’ diagnostic accuracy does not improve with training because feedback about incorrect psychological diagnosis is often absent, ambiguous, or highly delayed.

In spite of the mixed and seemingly negative results obtained for the impact of education and experience on diagnostic accuracy, we believe it is important to pursue this question empirically using carefully constructed groups differing in both education and experience. By doing so, we should be positioned to make a statement about the clinical implications of the effectiveness of training and who may be most likely to identify problematic behavior. Parents perceive teachers to be trustworthy experts in their child’s behavior in large part due to their extensive experience with children and their specialized educational training, so it is important to understand if insight is truly gained based on level of education or experience. In addition, previous studies investigating psychological diagnostic accuracy have had limitations that may have restricted their ability to find differences, and the influence of education and experience on diagnostic accuracy of children with developmental disorders has not been investigated. In the present study, using a brief play assessment, we examined the effects of education and experience in diagnostic accuracy of raters who evaluated a diverse sample of preschoolers. The effect of education and experience in categorization success generally (typical or atypical) and specifically (typical, SLI, ADHD, ASD) were analyzed, and main effects of education and experience were hypothesized. In addition, misclassification patterns were analyzed, where novices were suspected to categorize children as typical more often than experts. Finally, confidence in diagnosis was analyzed, with experts hypothesized to be more confident than novices. To the best of our
knowledge, no study to date has attempted to experimentally examine the influence of education and experience on adult raters’ ability to diagnostically classify preschool children based on their play performance.

Methods

Raters

In total, 44 raters were recruited based on level of education and experience, with groups formed based on the information provided in a self-report survey. Two education levels were recruited (low: novices, n=23; high: experts, n=21). Novices were either enrolled in a Bachelor’s degree program (freshman: n=5, sophomore n=4, junior: n=1, senior: n=3), or had completed a Bachelor’s degree but had not begun graduate level work (n=10). Experts were currently enrolled in a doctoral program (first year: n=2, second year: n=4, third year: n=3, fourth year: n=3, and fifth year: n=4) or had completed a terminal master’s degree (n=5). Novices and experts were further differentiated by experience level (low: inexperienced, n=21; high: experienced, n=23). Inexperienced raters had less than three years of experience working with young children (no experience: n=11, less than year of experience: n=7, between one and two years of experience: n=3), and all of those experiences held were on a very limited, part-time, volunteer basis, with none specific to preschoolers. Experienced raters had three to four years (n=19) or four to five years (n=4) of experience, although full or part time status varied. Experienced raters served as a daycare worker (n=1), ABA or behavioral therapist (n=5), nurse (n=1), teacher (n=5), classroom assistant (n=2), coach (n=1), master’s level clinician (n=3), or psychological clinician trainee (n=5). Four groups of raters comprise
the final sample: inexperienced novices (n=11), inexperienced experts (n=10), experienced novices (n=12), and experienced experts (n=11).

Materials

Child Experiences Questionnaire

The Child Experiences Questionnaire (CEQ) is a self-report survey designed for the present study to collect information about the rater’s education and past experiences with children with typical development and developmental disorders. Education background questions include: highest degree attained, current degree enrollment, and current degree year. Experiences questions include reporting of experiences with children with typical development and developmental disorders. Raters were first asked to list experiences with children with typical development across five specific domains: babysitting, camp counseling, clinical work, caring for neighbors, and caring for relatives. For each domain, the rater ranked their own level of experience from 0 to 2, with 0 indicating no experience, 1 indicating some experience, and 2 indicating a lot of experience. Raters were then asked to list experiences with children with developmental disorders across the same five specific domains on the same Likert scale, creating a total of 10 responses, ranging from 0 to 2 each. Raters were also encouraged to report on experiences with children that were not previously queried, within a free response format. Lastly, each rater was asked to list characteristics that exemplify preschoolers with typical development, SLI, ADHD, and ASD. During completion of this section of the survey, raters were instructed not to consult outside resources.

Responses from the CEQ were used to confirm group membership. Responses to the ten specific experience domains (five in typical and five in developmental disorders)
were totaled, creating a possible range of 0 to 20. Responses above seven confirmed experienced classification, while responses of seven or below confirmed inexperienced classification. Three raters originally classified as inexperienced fell above the threshold, indicating that they may be experienced. Free responses revealed that two of the three raters had more than three years of experience working with children, so they were subsequently moved to the experienced category. Six raters originally classified as experienced fell below the threshold, but all remained in the experienced category because they had three years or greater of experience working with children; four raters were teachers, and two were master’s level clinicians. In confirmation of group differences, experienced and inexperienced raters differed on CEQ total (inexperienced: range = 0-8, M= 3.81, SD= 2.21; experienced: range = 2-14, M= 8.57, SD=3.0), and no experience differences were noted between education groups (novice: range = 0-14, M= 6.30, SD=3.80; experts: range = 2-14, M=6.29, SD=3.33).

Play Videos

Raters watched 93 videos of children completing a standard free play task (Affect in Play Scale – Preschool Version (APS-P; (Kaugars & Russ, 2009). The APS-P is a semi-structured play task in which the child interacts with the examiner for a short period of time before completing five minutes of independent free play. During administration, the child is seated at a small table across from an examiner. The child is positioned so that both the child and the toys on the table are fully visible during the videotape recording.

In this study, the semi-structured task began with the examiner stating that she was there to learn how children play. Then, she removed the toys one by one from her
bag, placing them on the table in front of the child, requesting that the child label each
toy and assisting when necessary. The bag of toys included small stuffed animals (large
elephant, dog, cat, alligator, polar bear), small plastic animals (cheetah, small elephant,
zebra, shark, giraffe) and other miscellaneous toys (race car, three cups of different sizes,
and a ball). Once all of the toys were presented and verbally labeled, the examiner
modeled a story for the child. The script was as follows: “Now we’re going to make up a
story using the toys on the table. See how you can play with the toys. This is the bear. He
says, ‘I’m really hungry! Where can I find some food? Oh look, I found some cookies. I
love cookies. Yum! Yum! Here’s another cup. Oh yucky! I don’t like what’s inside there!
Yuck!’” The child was then told to make up a story while playing with the toys as the
examiner did her homework. The child was prompted to talk out loud while playing to
better assist the examiner in her understanding of the child’s play script. Throughout the
child’s time in independent play, the examiner actively monitored the child’s progress
and appropriately redirected behavior. Redirections were given when needed to
encourage the child to complete the task, but care was taken by the examiner not to
directly assist in scaffolding play behavior.

Each video was viewed from the beginning to the end of the play period, usually 5
minutes in length. Each play video varied somewhat in length from 5 min 20 s to 8 min
30 s (M=7 min, SD=30 s). The length of time for set-up varied due to administration
demands.

The play videos were collected as a part of a larger study investigating play
differences in children with developmental disorders. Children in the videos ranged in
age from four to seven (M=5.70, SD=0.98). Children with atypical development (n=70)
and children with typical development (n=23) were included. Children with atypical
development were previously diagnosed with a developmental disorder by a licensed
psychiatrist, neurologist, or speech language pathologists (SLI: n=24, ADHD: n=26,
ASD: n=20) and were highly functioning. Children had performance IQs that ranged
from 73 to 142 on the Wechsler Preschool and Primary Scale of Intelligence- Third
Edition (WPPSI-III; Wechsler, 2002), with diagnostic groups not different on
performance IQ.

Rating Form

A rating form was created for the purposes of this study to collect diagnostic
information from the adult raters on each of the child participants. The rating form
contained general information followed by diagnostic information. General information
included the rater’s name, child’s number, and the child’s first name. Diagnostic
information was separated into two categories of ratings: general and specific. In the
general portion of the diagnostic rating form, the raters indicated the following
information about each of the 93 child play tapes: typical or atypical development,
confidence level in the general classification, and the reasons underlying classification.
The confidence level was rated on a Likert scale (1: Not Confident, 2: Somewhat
Confident, 3: Confident, 4: Very Confident).

The specific section of the diagnostic rating form was only completed if the rater
in the general section indicated atypical development. If the rater indicated the participant
was atypical, they were forced to choose from one of three diagnostic categories: SLI,
ADHD, or ASD. Consistent with the previous rating, confidence level in the specific
classification was requested, as well as the reasons underlying the classification chosen.
The confidence level was rated on the same Likert scale (1: Not Confident, 2: Somewhat Confident, 3: Confident, 4: Very Confident).

Procedure

Following consent into the study, raters completed a 45-minute one-on-one training session where they completed the CEQ, received training on the procedure for watching the videos independently, and completed a practice video. As part of the informed consent process, raters were first asked to agree not to consult outside resources or other raters in the study to assist in rating the play videos. In addition, raters were told that they could only watch each of the videos one time.

Viewing of the play videos was conducted independently by each of the raters to allow for flexibility in scheduling and to ensure individual effort rather than collaboration. During independent viewing, raters were instructed to watch videos for only one to two hours at a time in order to prevent potential fatigue. Independent ratings were collected over a three-week time period, requiring participation of approximately five hours per week to complete the 93 tapes.

During independent viewing, all raters followed the same predetermined order for watching the play videos. After the rater watched the video, he or she filled out the electronic rating form. Given the educational nature of this project, all raters were provided a supplemental educational summary in poster format after the conclusion of data collection.

Data Analysis

Continuous dependent variables were analyzed using univariate or multivariate analysis of variance. Levene’s test of equality of error variances was used to confirm the assumption that variances among groups were equal, and data were transformed in cases
of violation (Field, 2013). In addition, Mauchly’s test of sphericity was used to assess the sphericity assumption when there were three levels of the independent variable. When the sphericity assumption was violated, Greenhouse-Geisser was used to estimate degrees of freedom (Greenhouse & Geisser, 1959).

Results

In order to address how education and experience influences raters’ ability to identify typical development, an analysis of variance (ANOVA) was conducted. Education level (0,1) and experience level (0,1) were the between subject variables, and typical accuracy (rate of children correctly classified as typical) was the dependent variable. Levene’s test of equality of error variances was significant, $F(3,40)=3.35$, $p = .028$, so a reciprocal transformation was performed. No significant main effects emerged for education, $F(1, 40)=.047$, $p=.83$, partial $\eta^2 = .001$ or experience, $F(1, 40)=.63$, $p=.43$, partial $\eta^2 = .015$ (see Table 1). In addition, there was no interaction between experience and education, $F(1, 40)=.13$, $p=.72$, partial $\eta^2 = .003$. A second ANOVA was conducted to assess the impact of education and experience on diagnostic accuracy in identifying atypical development. Again, no significant main effects for education, $F(1, 40)=.66$, $p=.42$, partial $\eta^2 = .016$ or experience emerged for atypical diagnostic accuracy, $F(1, 40)=.47$, $p=.50$, partial $\eta^2 = .012$ (see Table 1). There was also no interaction between education and experience on atypical diagnostic accuracy, $F(1, 40)=.010$, $p=.92$, partial $\eta^2 <.001$.

Although no effect was obtained for education and experience on atypical or typical diagnostic accuracy, specific diagnostic accuracy (typical SLI, ADHD, ASD) was investigated for effects of education and experience next. Percentages of correctly classified children in each specific diagnostic category (typical, SLI, ADHD, ASD) were
the dependent variables in a 2x2x4 ANOVA, with education (0,1) and experience (0,1) as the between subjects variables and diagnosis (typical, SLI, ADHD, ASD) as the within subjects variables. Mauchly’s test of sphericity indicated a violation of the sphericity assumption $X^2(5) = 18.14$, $p = .003$, therefore Greenhouse-Geisser ($\varepsilon = .82$) was used to estimate degrees of freedom. Levene’s test $F(3,40) = 3.35$, $p = .028$ was also violated for typical accuracy, so a reciprocal transformation was performed on all independent variables. No main effect of education, $F(1, 40) = 3.79$, $p = .060$, partial $\eta^2 = .087$ or experience, $F(1, 40) = 1.48$, $p = .23$, partial $\eta^2 = .036$ emerged for diagnostic accuracy. In addition, no interactions between diagnosis, experience, and education reached conventional levels of significance. However, a main effect for diagnosis type emerged for specific diagnostic accuracy, $F(2.47, 98.90) = 36.00$, $p < .001$, partial $\eta^2 = .47$ (see Table 2). Post-hoc tests with Bonferroni correction for multiple comparisons indicated that children with typical development were easier to diagnose than children with SLI, ADHD, or ASD, $p < .001$. In addition, children with ASD were easier to diagnose than children with ADHD, $p = .002$ or SLI, $p = .056$. There was no difference in ease of diagnosis between children with SLI and children with ADHD, $p = 1.00$.

Next, misclassification trends were analyzed for effects of education and experience using four separate 2x2x3 ANOVAs (one for each diagnostic category). For example, the ANOVA for misdiagnosis of children with SLI included misdiagnosis rates in the three other specific categories (typical, ADHD, ASD). Education (0,1) and experience (0,1) were the between subjects variables, and percentage of misclassified children in each specific category were the within subjects variables. There were no main effects for education, maximum $F(1,40) = 2.70$, $p = .11$, partial $\eta^2 = .063$, or experience,
maximum $F(1,40)=3.40$, $p=.073$, partial $\eta^2 = .078$ (see Table 3). However, there were significant main effects for type of misdiagnosis for all three of the separate developmental disorder analyses, minimum $F(1.85,73.98)=22.03$, $p<.001$, partial $\eta^2 = .36$ but not on the typical ANOVA $F(1, 40)=1.43$, $p=.24$, partial $\eta^2 = .034$ (see Tables 3, 4). Post hoc tests with Bonferroni correction revealed that in all cases, children with developmental disorders were most likely to be misclassified as having typical development, $p<.001$. No differential preference was demonstrated for any of the atypical diagnostic categories, $p=1.00$.

Next, an analysis was conducted to investigate if education or experience affects ability to discriminate between typical and atypical development. Signal detection theory was used in the creation of a single unit to represent accuracy, $d'$ (MacMillan, 2002). Accuracy ($d'$) was calculated for each subject by subtracting standardized false alarm frequency (number of children with typical development falsely identified as having atypical development divided by the total number of children with typical development) from the standardized hit rate frequency (number of children with atypical development correctly identified as having atypical development divided by the total number of children with atypical development). An analysis of variance (ANOVA) was conducted, with education level (0,1) and experience level (0,1) as the independent variables and accuracy ($d'$) as the dependent variable. No significant main effect of education, $F(1,40)=.006$, $p=.94$, partial $\eta^2 < .001$ or experience, $F(1, 40)=.76$, $p=.39$, partial $\eta^2 = .019$ emerged for $d'$ (see Table 1). There was also no interaction between education and experience on $d'$, $F(1, 40)=.51$, $p=.48$, partial $\eta^2 = .012$. Thus, ability to discriminate between atypical and typical development was not related to education or experience.
Lastly, an ANOVA was conducted to determine if education, experience, or accuracy were related to confidence. Confidence served as the dependent variable, with education (0,1) and experience (0,1) as the between subjects variables, and correctness classification (0: wrong, 1: correct) as the within subjects variable. Again, there was no main effect of education, $F(1,40)=.016$, $p=.90$, partial $\eta^2 < .001$ or experience $F(1,40)=2.08$, $p=.16$, partial $\eta^2 = .049$ on confidence, nor were any of the interactions significant. However, there was an effect of accuracy rating on confidence level, Wilk’s $\lambda = 0.62$, $F(1,40)=24.70$, $p<.001$, partial $\eta^2 = .38$, with all raters more confident when they were correct ($M=2.56$, $SD=.44$) than when they were incorrect ($M=2.39$, $SD=.43$).

Discussion

Contrary to expectation, education and experience level in the present study did not predict ability to accurately diagnose children with typical development, atypical development, or any of the three developmental disorders (SLI, ADHD, ASD). In addition, neither education nor experience level predicted the pattern of misdiagnosis, success in discriminating between typical and atypical development, and confidence in diagnosis. Overall, it appears that neither education nor experience reliably predicts the skill of these raters in identifying the developmental or diagnostic category of preschool children engaged in a brief free play activity.

Although there were no main effects of education or experience, information about trends for diagnosis in play were achieved. According to signal detection theory, discriminability greater than zero indicates an above chance skill in diagnosis (MacMillan, 2002), and all raters had discriminability greater than zero. Therefore, all 44 raters demonstrated an above chance skill in discerning between typical and atypical development. Thus, the use of a brief play tape appears to provide enough information
that people can reliably discriminate between typical and atypical development even though education and experience did not reliably predict improvements in the task. This finding is an exciting affirmation that play may be a useful tool to include in assessment of children, as it provides an inexpensive, brief, non-threatening environment to gather developmental information (Short, Noeder, Gorovoy, Manos, & Lewis, 2011).

Results also revealed interesting individual differences, with some diagnostic classifications (typical and ASD) easier to identify than others (ADHD and SLI). All children, including children with ASD, were high functioning, and performance intelligence was comparable across groups. Thus, intellectual functioning differences do not account for the differential success in identifying children with ASD or typical development. Although the reason for higher rates of ASD and typical development accuracy remains unclear, we hypothesize that the absence of atypical behavior (i.e., compromised pretend play, inability to narrate play, hyperactivity) may be responsible for the high levels of accuracy in typical development identification. In contrast, multiplicity of impairments (i.e., compromised language and behavior) present in children with ASD may have enabled recognition of diagnosis more readily than children with only one domain of impairment (SLI: language; ADHD: behavior). Raters may have been reticent to categorize children in an atypical developmental category if they showed symptomatology in just one area but felt more comfortable diagnosing when the child appeared to be impaired in multiple domains.

Furthermore, misclassification patterns are also consistent with the explanation that raters may have been reticent to diagnose atypical development when a child was impaired in one domain. Children in the developmental disorder groups were most often
misclassified as having typical development over the other two disordered groups. When classifying children with typical development, conversely, raters did not show preference to any one of the three developmental disorder classifications. Therefore, decision-making appears to occur more often between a developmental disorder classification and typical development rather than between two developmental disorder categories.

In addition, confidence appears to change based on correctness of diagnosis, in spite of lack of differences in education and experience. Results are consistent with other categorization tasks such as eyewitness identification, where confidence is somewhat related to accuracy (Sporer, Penrod, Read, & Cutler, 1995). Confidence has also been associated with the amount of supporting evidence the participant can provide in their response (Koriat, Lichtenstein, & Fischhoff, 1980). Perhaps raters were more likely to be correct when they could provide more evidence in favor of their diagnosis, thus increasing their level of confidence.

Although diagnostic information can be gleaned from the present study, the suspected main effects of education and experience were not present. The insignificant results of education and experience on assessment of preschoolers is consistent with findings obtained for classification of adults with brain injury (Goldberg, 1959) and psychopathology (Aarts et al., 2012; Witteman & van den Bercken, 2007) by adults with differing levels of expertise.

Dawes, in his book *House of Cards* (1994), makes the argument that the reason training has not been shown to improve diagnostic accuracy in psychologists is in part because they do not undergo routine practice with direct, consistent feedback about their accuracy in diagnosis. Thus, he argues, as people gain more experience, their schemas do
not improve because they do not receive feedback. Deliberate practice makes a similar argument, where expertise is only achieved when people are motivated toward a specific goal and practice and feedback are directly targeted toward achieving that goal (Ericsson et al., 1993).

Given the fact that education and experience levels were not identified as important predictors in diagnostic ability in the current investigation, should we conclude that these factors indeed do not matter and launch a campaign designed to unearth the importance of specialized training and experiential learning? For a variety of reasons, it may be a mistake at present to argue this position. Limitations in the present study include restricted range of education, limited range of experience, mismatch between type of experience and testing, and homogeneity of group demographics.

Limitations

There are many reasons why education and experience are not shown to reliably predict the improvement of diagnostic ability in the present design. First, the participants included represent a limited range in education, and the spread between the two levels of education (between having a Bachelor’s degree and being a first year graduate student) may not have been wide enough to capture the potential differences that education may have on diagnostic performance. In addition, very little of graduate training is devoted to diagnostic assessment of preschoolers (one week of a six week module). Educational differences while present, but the education received by the experts was not specifically tailored to preschool diagnosis. Diagnostic experts in the field improve on their educational experiences by participating in supplementary workshops and educational training on an ongoing basis after graduation. For this reason, our so-called experts do not represent the range of education that may be had by more senior experts.
Second, the experience of the raters does not appear to meet the requirements of deliberate practice, as experienced groups were not involved directly in diagnosis and feedback for their 3-5 years of experience. Instead, experience was determined by exposure to children, and experienced raters had a wide range of roles when interacting with children (i.e., daycare worker, behavioral therapist, nurse, teacher, classroom assistant, coach, clinician, psychological clinician trainee). In addition, the positions held by the coach, ABA or behavioral therapists, nurse, and training psychological clinicians were only held on a part time basis. Apart from some practice in general diagnosis for master’s level clinicians and students training in clinical child psychology, other members of the experienced group likely did not obtain direct practice in diagnosis and feedback on accuracy. In addition, few of the clinicians received direct experience in the diagnosis of preschool children. Even for raters who did have practice in diagnosis, the majority of the work by master’s level clinicians and training psychological clinicians is not solely devoted to time spent on time diagnosing, as therapy, case evaluation, and assessment are other components of the job. Within the deliberate practice framework, people who establish expertise in an area must undergo practice with consistent, unambiguous and direct feedback for modification of goal-directed performance (Ericsson et al., 1993). Although supervision is a required form of monitoring in clinical psychology, students do not receive the type of immediate, unambiguous feedback provided by other fields such as sports, music, and games that have higher rates of relation between deliberate practice and performance. It remains to be seen whether a student who had undergone true deliberate practice in the field of psychological assessment would perform at a higher accuracy level than people in the present study.
Third, the experience of raters represented a limited time range, with the maximum amount of experience at only five years, much shorter than the ten-year rule for establishing expertise (Simon & Chase, 1973). In addition, few had direct experience diagnosing preschool children. The difference between the amount of experience evidenced by the inexperienced group and the experienced group may not have been distinct enough to result in differences. Additionally, the present sample does not include adults with higher levels of experience, and therefore it remains unclear how experts with experience past five years would perform.

Fourth, the characteristics of the raters themselves may have minimized ability to achieve differences across groups. Raters were all of similar background and intelligence level. All raters attended competitive midwestern universities. In addition, groups represented a very small sample, with only 10-12 raters per group.

Fifth, preschoolers overall are very difficult to diagnose because symptoms of SLI, ADHD, and ASD overlap tremendously. Adults accept a wide latitude of behavioral and language performance as the norm at this age due to the individual differences in development and the opportunity for greater maturation with time. Therefore, the difficulty of the task may have limited the ability to find differences.

Sixth, raters agreed not to seek outside sources, discuss their responses with other raters, or repeat videos. Due to the extensive amount of time the task required, there was no way to guarantee that the raters followed all of these rules. If groups acted differently, for example if undergraduate students more commonly violated guidelines than other groups, then that may have limited our ability to find differences.
Clinical Implications

Despite the lack of significant findings related to the importance of education and experience for diagnostic accuracy, clinical implications are still clear from the present study. First, the finding that all of the participants were able to reliably differentiate between children with typical and atypical development strongly supports the efficacy that a brief play assessment contains enough information about preschoolers to allow raters of limited education and experience to reliably predict group membership. Children engaged in play are able to display their skills in a non-threatening environment on an inexpensive and brief task that has the potential to provide valuable information for diagnostic categorization (Short et al., 2011). Perhaps future studies could investigate whether or not targeted training, both in providing diagnostic information and deliberate practice, would change accuracy levels. If targeted training does change accuracy levels, opportunities for training among important individuals close to children (i.e., parents, teachers, daycare workers) should be considered.

In addition, the results indicate that people are better able to identify children with ASD, perhaps because of multiple domains of impairment, than they are to identify children with single domains of impairment like those seen in ADHD or SLI. We speculate that raters were reticent to assign a diagnosis without witnessing multiple domains of impairment and were more cautious when only one domain of impairment was evidenced. This cautious strategy while appropriate with young children who often display wide latitude of acceptable behaviors may be less than ideal when early intervention and prevention is the more appropriate course of action for children who present with developmental disorders. Nonetheless, the findings from the present study
are exciting in that they provide evidence for the utility of a brief play assessment in identifying children who may be in need of services.
Table 1: General diagnostic accuracy

<table>
<thead>
<tr>
<th></th>
<th>Novices Inexperienced</th>
<th>Novices Experienced</th>
<th>Experts Inexperienced</th>
<th>Experts Experienced</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atypical accuracy</td>
<td>.58 (.21)</td>
<td>.61 (.16)</td>
<td>.62 (.13)</td>
<td>.65 (.14)</td>
</tr>
<tr>
<td>Typical accuracy</td>
<td>.66 (.24)</td>
<td>.62 (.16)</td>
<td>.65 (.13)</td>
<td>.59 (.20)</td>
</tr>
<tr>
<td>d’</td>
<td>.78 (.37)</td>
<td>.63 (.30)</td>
<td>.72 (.32)</td>
<td>.70 (.30)</td>
</tr>
</tbody>
</table>
Table 2: Specific diagnostic accuracy

<table>
<thead>
<tr>
<th></th>
<th>Novices</th>
<th>Experts</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Inexperienced</td>
<td>Experienced</td>
<td>Inexperienced</td>
<td>Experienced</td>
</tr>
<tr>
<td>Typical</td>
<td>.66 (.24)</td>
<td>.62 (.16)</td>
<td>.65 (.13)</td>
<td>.59 (.20)</td>
</tr>
<tr>
<td>SLI</td>
<td>.21 (.18)</td>
<td>.32 (.16)</td>
<td>.32 (.13)</td>
<td>.32 (.13)</td>
</tr>
<tr>
<td>ADHD</td>
<td>.25 (.12)</td>
<td>.31 (.14)</td>
<td>.30 (.097)</td>
<td>.35 (.082)</td>
</tr>
<tr>
<td>ASD</td>
<td>.37 (.15)</td>
<td>.38 (.14)</td>
<td>.40 (.13)</td>
<td>.36 (.16)</td>
</tr>
</tbody>
</table>
Table 3: Rates of correct and incorrect classification

<table>
<thead>
<tr>
<th>True Diagnosis</th>
<th>Typical</th>
<th>SLI</th>
<th>ADHD</th>
<th>ASD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical</td>
<td>.63 (.18)</td>
<td>.12 (.10)</td>
<td>.15 (.10)</td>
<td>.12 (.10)</td>
</tr>
<tr>
<td>SLI</td>
<td>.41 (.19)</td>
<td>.29 (.15)</td>
<td>.15 (.089)</td>
<td>.15 (.12)</td>
</tr>
<tr>
<td>ADHD</td>
<td>.41 (.17)</td>
<td>.14 (.084)</td>
<td>.30 (.11)</td>
<td>.14 (.10)</td>
</tr>
<tr>
<td>ASD</td>
<td>.32 (.17)</td>
<td>.14 (.10)</td>
<td>.15 (.10)</td>
<td>.38 (.14)</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Misclassification</td>
<td>Education</td>
<td>Experience</td>
<td></td>
</tr>
<tr>
<td>-----------</td>
<td>------------------</td>
<td>-----------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>SLI</td>
<td>$F(1.30, 51.90)=38.93$</td>
<td>$\eta_p^2=1.22$</td>
<td>$\eta_p^2=1.34$</td>
<td></td>
</tr>
<tr>
<td>ADHD</td>
<td>$F(1.41, 56.22)=46.72$</td>
<td>$\eta_p^2=1.78$</td>
<td>$\eta_p^2=2.67$</td>
<td></td>
</tr>
<tr>
<td>ASD</td>
<td>$F(1.85, 73.98)=22.03$</td>
<td>$\eta_p^2=.016$</td>
<td>$\eta_p^2=0.089$</td>
<td></td>
</tr>
<tr>
<td>Typical</td>
<td>$F(1, 40)=1.43$</td>
<td>$\eta_p^2=.36$</td>
<td>$\eta_p^2=.078$</td>
<td></td>
</tr>
</tbody>
</table>
References


